# Innovating Health Technology Assessment Methods: Barriers and Enablers illustrated using Qualitative and Quantitative Methods on Real-World Data

Li Jiu

The research presented in this thesis was performed at the division of Pharmacoepidemiology and Clinical Pharmacology of the Utrecht Institute for Pharmaceutical Sciences (UIPS), Faculty of Science, Utrecht University, Utrecht, the Netherlands.

**Cover:** Li Jiu; Lin Li
**Layout & Printing:** Proefschriftenprinten.nl, Ede
**DOI:** https://doi.org/10.33540/1310
**ISBN:** 978-90-834024-7-5

Financial support for printing of this thesis was kindly provided by the Utrecht Institute for Pharmaceutical Science (UIPS). This PhD project was funded by the HTx project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825162.

# Innovating Health Technology Assessment Methods: Barriers and Enablers illustrated using Qualitative and Quantitative Methods on Real-World Data

**Innovatieve beoordelingsmethoden voor pakketbeheer: uitdagingen en kansen bij de toepassing geïllustreerd op basis van het gebruik van kwalitatieve en kwantitatieve methoden om klinische praktijkgegevens te analyseren**

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
donderdag 28 maart 2024 des ochtends te 10.15 uur

door

**Li Jiu**

geboren op 8 februari 1994
te Jiangsu, China

**Promotoren:**
Prof. dr. W.G. Goettsch
Prof. dr. A.K. Mantel - Teeuwisse

**Copromotor:**
Dr. J. Wang

**Beoordelingscommissie:**
Prof. dr. A. de Boer
Prof. dr. D.M.J. Delnoij
Dr. T.L. Feenstra
Dr. H. Gardarsdottir
Prof. dr. L. Hooft

# Table of contents

Chapter 1

# General Introduction

## Health technology assessment methods: definition and categorizations

Health technology assessment (HTA) is a process of using explicit methods to determine the value of a health technology at different points in its lifecycle (1). With HTA, stakeholders, such as patients, clinicians, industry, HTA agencies, and researchers, can be better informed to support decision-making on reimbursement and pricing or decision-making in clinical practice (2,3). HTA may add great value to the whole human society, as it aims to contribute to an efficient and equitable health system. On the one hand, it could ensure that the health outcomes of patients and individuals could be improved as much as possible, with the limited healthcare resources (4). On the other hand, HTA may ensure that patients who urgently need healthcare have access to timely treatment, while patients with minor health conditions could avoid treatment which incurs unnecessary costs (4,5).

While HTA is valuable to the healthcare system, the quality and relevance of HTA methods is often discussed, for instance on the question of whether HTA methods are appropriate. For example, novel health technologies, such as digital health and machine learning technologies, involve features (e.g. continuously updated algorithms and new ethical challenges) that may need special considerations during the HTA process, while the existing HTA methods may not structurally take these features into account (6,7). Another example is that, as the organization of healthcare varies across countries, the application of an HTA method, originally developed in a certain setting to another setting, can be questioned (8). Consequently, stakeholders could make sub-optimal decisions, e.g., on reimbursement and pricing, based on evidence obtained and synthesized with inappropriate HTA methods. To increase the availability of appropriate HTA methods, HTA methods have been repeatedly developed and implemented, since HTA became an important element of healthcare systems in the 1980s (9,10).

"HTA methods" is an umbrella concept with broad implications, but without a consistent definition. Still, the concept "HTA methods", or its synonyms (e.g. HTA tools), has occurred frequently in HTA agencies which provide methodological guidance on high-quality HTA or research projects that focus on methodological research. The European network for health technology assessment (EUnetHTA) has developed the HTA Core Model to facilitate production and sharing of HTA information, and to inform decision-making (11-13), see Figure 1. The HTA Core Model categorizes HTA value into nine domains, such as safety, clinical effectiveness, cost and economic effectiveness, ethical analysis, legal aspect, etc. For each HTA quality concern within a domain, the HTA Core model illustrates the methods suitable for assessing the HTA

quality. For example, one concern in "the cost and economic effectiveness" domain is "the uncertainties surrounding the costs and economic evaluation(s) of the technology and its comparator(s)". Correspondingly, the HTA Core model recommends the use of a deterministic sensitivity analysis in tabular form or using a Tornado diagram. Also, one concern in the "clinical effectiveness" domain is to assess all benefits and harms of a technology, including but not limited to mortality and quality of life. To address the concern, the HTA Core Model suggests applying methods to integrate trials, observational studies, and modelling studies. Country-specific HTA agencies have also published guidelines for conducing health economics analysis, such as the "guideline for economic evaluations in healthcare" published by the Dutch National Institute of Health Care (ZIN) in 2016 and the manual of HTA evaluations published by the National Institute for Health and Care Excellence (NICE) in 2022 (14,15). To provide further guidance on HTA, HTA agencies have also published a series of methodological guidance. For example, ZIN has proposed the guideline for building cost-effectiveness models in R (16), which introduces methods for structuring a model or testing model validity. ZIN has also proposed a guideline for outcomes research, which illustrates methods to collect or analyze different types of data, such as costs, patient-reported outcomes, data from trials, observational studies, or cross-sectional registries, etc. (17). Similarly, NICE has published reports to recommend methods for collecting and synthesizing HTA evidence (e.g. future unrelated health costs), for structured decision-making (e.g. equality issues), and for assessing value of challenging technologies (e.g. histology independent cancer drugs) (18).

In addition to HTA agencies and research projects, scientific articles that provided suggestions on how to improve HTA quality in regions or countries also have offered a way of describing "HTA methods". For example, in the United Kingdom (UK), HTA decision makers have relied on more uniform evidence appraisal methods (e.g. from NICE) in their HTA process, so method transparency could be somewhat guaranteed (4). Additionally, as mentioned by Diego et al. in 2017 (19), "there can be several tools and methods that could improve the quality of HTA implementation". Scientific literature also covers "HTA methods" extensively. For example, regarding appraisal of HTA evidence, methodological reviews have been published to summarize methods used to synthesis evidence from different sources (20), methods used to identify errors in health economics models (21), or methods for conducting budget impact analyses (22). Regarding HTA decision-making, reviews have been published to summarize methods for taking into account multiple decision-making criteria (i.e. multi-criteria decision analysis) (23) and methods for making decisions in limited timeframes (i.e. rapid review) (24). In addition, some reviews have focused on methodological issues for a certain HTA domain, such as methods to address ethical, legal, or organizational aspects (25-29).

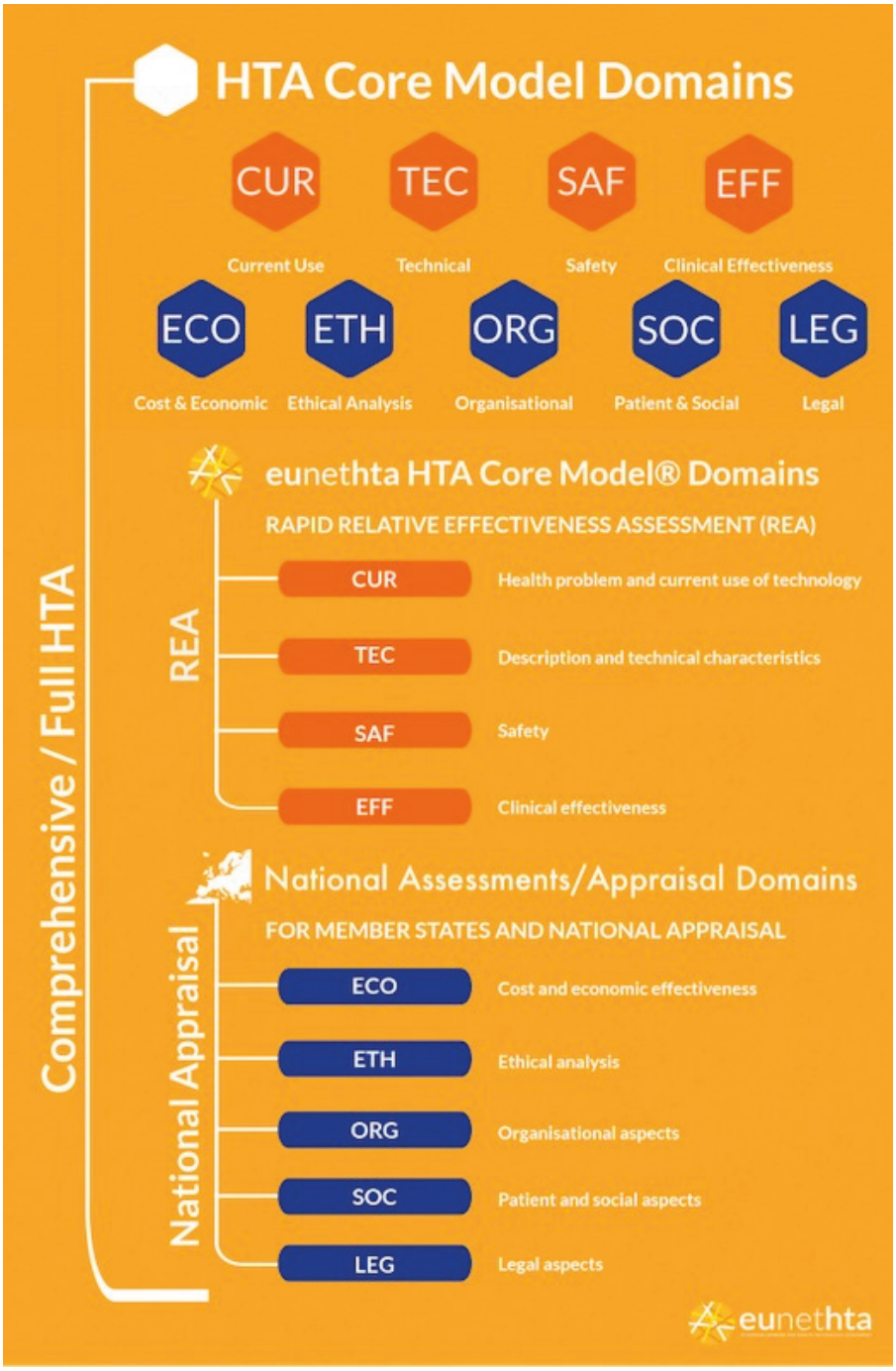



**Figure 1. Domains of HTA value defined by the HTA Core Model (adapted from (11)).**

According to the way "HTA methods" are mentioned by HTA agencies and in research projects and scientific articles, we can see that the concept "HTA methods" has several features. First, the concept embraces the full scope of an HTA process, including evidence collection, evidence appraisal, decision-making, and monitoring (30). Second, HTA methods can be divided into multiple categories. They can be qualitative, such as frameworks, guidelines, and checklists, or quantitative, such as models and statistical approaches. Third, HTA methods, regardless of their functions or categories, involve a phase of development or implementation. Usually, an HTA method is first developed (e.g. in a research project), then disseminated (e.g. through publication), and finally applied in a HTA setting.

## HTA method development and implementation: the necessity and general problems

Since HTA became an important element of healthcare systems in the 1980s, HTA methods have been developed and implemented repeatedly (9,10). Development indicates a process in which an HTA method becomes more advanced (31), while implementation indicates the act of starting to use an HTA method (32). While HTA methods appear in large numbers, some general problems that have negatively affected method development or implementation have occurred over time. To clearly illustrate these problems, in the following paragraphs general problems on method development and implementation will be discussed using RWD-related methods as examples.

### Problems related to HTA method development.

One type of problem related to developing HTA methods is the lack of a clear overview of the needs from HTA stakeholders. For example, hundreds of risk prediction models have been developed to provide prognosis information on the occurrence of a disease or disease complication (33-35). These risk prediction models mainly function as tools for clinical decision-making (36,37). While they can also be used as a part of a health economics model, this function is normally not recognized by the model developers (36). Consequently, these models often lack technical features, such as predicting the probability of disease occurrence in a one-year time cycle, that could make them fit well into a health economics model. Although the importance of understanding the needs from HTA stakeholders has been increasingly recognized in the HTA field, the approaches that facilitate the understanding from method developers are still lacking.

Another type of problem, related to HTA method development, is the limitation of resources, such as available time, high-quality data, and knowledge across research

disciplines. For example, it is well recognized that the lack of high quality data is often a barrier to HTA method development, especially when a method needs to be externally validated in various settings (38). Another problem is the lack of theory for understanding how to develop HTA methods. As mentioned by Shenhar et al. in 2007 and 2016, as no single comprehensive framework for understanding innovation challenges in highly complex research exists, the similarities and differences of research that involve complex innovation activities should be further explored by stakeholders (39,40). In the HTA context, the lack of conceptual research may cause misconception on similarities or differences of a process of developing an HTA method, thus reducing efficiency in method innovation.

### Problems related to HTA method implementation

One type of problem related to implementing HTA methods is the lack of expertise on the HTA methods. For example, to assess quality of primary studies investigating efficacy of a healthcare intervention, the lack of knowledge on how to use the appraisal tools (e.g. ROBINS-I) is often a barrier of implementing the tools (41). The variety of appraisal tools, each of which involve some specific knowledge on how to use the tool, could even further complicate the problem of implementing the tools (42,43). Similarly, statistical approaches that merge different types of data, such as those from randomized clinical trials (RCTs) and daily practice ('real-world data', RWD) in a (network) meta-analysis are presented in complex math formulars, and can only be used with a specific statistical software (e.g. WinBUGS) (44,45), which a HTA stakeholder may not be familiar with.

Another problem of method implementation is the lack of collaboration skills that enable the use of different research methods. For example, in the case of implementing a health economics model which utilizes RWD and patient experiences, engaging a large and diverse stakeholder group, including patients, clinicians, payers, and researchers, can increase the scope and complexity of model implementation. As mentioned by Xie et al., one challenge of engaging the HTA stakeholders is to use tailored communication strategies to address different research questions (e.g. general questions or questions that need specific expertise) (46). Additionally, Xie et al. emphasized the necessity of refining methodology to synthesize conflicting viewpoints or potentially missing stakeholder perspectives in the model application (46).

Due to the existence of general problems that may negatively affect method development or implementation, few new HTA methods have been applied, after they were developed. For example, according to Van Giessen et al. health economics evaluations of risk prediction models and studies investigating the model impact were

rare despite the huge number of risk prediction models in the medical literature (36). Also, according to Quigley et al. in 2019, while more than 40 tools that assessed quality of non-randomized studies had been developed, users still lacked consensus on how to select and use the tools (47).

One potential solution to the above-mentioned general problems is to establish and illustrate a pattern of identifying the stakeholders' needs and facilitating stakeholder collaboration, throughout the innovation process. Previous research has built some foundations, by developing guidance for developing some types of HTA methods. For example, several guidelines have been published for developing health economics models (48,49) and patient-reported outcome measures (50). However, these studies have some limitations. First, they only focused on method development but did not guide on how the HTA methods should be implemented or transferred to another therapeutical or geographical context. Second, these studies did not guide in understanding why a method should be developed. In other words, the ways of identifying the needs remained uncertain to HTA stakeholders. The need for an HTA method is worth investigating, as understanding the needs may substantially improve the method quality, e.g., in terms of transferability (51). Third, this previous research only focused on one or several types of HTA methods (e.g., collection of evidence from patients), and could not function as the general guidance for innovating all types of HTA methods. Although HTA methods vary greatly in format and function (e.g., qualitative and quantitative methods), they may have similarities in their pattern of innovation, which is worth investigating and understanding by all relevant stakeholders. Therefore, further research is needed to provide a general guidance on how HTA methods should be innovated and to compare the patterns of innovation among the different types of methods.

## The needs for methods to promote the use of real-world data in HTA settings

One of the issues that has been on the agenda of HTA agencies is the appropriateness of the use of the different types of data in HTA and the quality and consistency of its associated methods. Historically, HTA agencies, especially in the field of medicines, have relied on the use of randomized clinical trials because it provides as much as possible unbiased estimation of the relative effect of a (new) health technology to its comparator (52). However, for several reasons, such as changing remits of HTA agencies and the lack of RCT data for complex and personalized therapies, there is a growing appetite to also use other non-randomized clinical data in HTA (53,54).

Real-world data (RWD), broadly speaking, refers to data collected in a setting beyond RCTs (55,56), but their definitions could vary in settings. According to Makady et al., one of the RWD definitions is data collected without interference with treatment assignment (55). RWD has become increasingly popular in HTA, mainly because of the growing demand for evidence on effects of healthcare interventions collected beyond experimentation conditions (56). For example, estimation of cost-effectiveness with a modelling approach which relies on RWD has become a gold-standard method (57). Also, incorporation of RWD into a reassessment process of health technologies for updating a reimbursement decision has become a promising strategy (58). Generally speaking, RWD can complement RCTs, especially when RCTs are scarce or infeasible to conduct (59).

However, the usefulness of RWD is often questioned due to quality concerns. According to the Cochrane Handbook, RWD have higher risk of bias than RCTs, and are vulnerable to various types of bias, such as selection and confounding bias (60). Inclusion of such RWD without evaluating or addressing these concerns would lead to questionable conclusions on value (e.g., regarding effectiveness or cost-effectiveness) of a healthcare intervention. Consequently, RCTs are still often used as the only evidence for the purpose of HTA decision-making (61,62).

To facilitate the use of RWD, the quality concerns need to be evaluated and then addressed with appropriate methods. For example, to evaluate RWD quality, HTA stakeholders need an appraisal tool (e.g. Risk Of Bias In Non-randomized Studies - of Interventions (ROBINS-I)) for assessing various risk-of-bias domains (63). Then, the RWD with high-quality concerns should be discarded, while RWD with moderate concerns might need to be downweighed, if they are combined with RCTs as evidence, before used for the decision-making purpose. In the case of meta-analyses (MAs) or network meta-analyses (NMAs), a commonly used methodology for synthesizing HTA evidence, RWD can be downweighed with some statistical approaches, such as power prior (64). Considering the emergence of novel health technologies and the variety of settings where HTA is conducted, as mentioned in the earlier section, the existing methods for evaluating or addressing RWD quality concerns may not satisfy all needs of HTA stakeholders, so novel methods need to be repeatedly developed and implemented.

## Thesis Objective

In summary, understanding the definition and categorizations of HTA methods may help improve HTA quality and facilitate multi-disciplinary collaboration within the HTA context. In addition, HTA methods need to be repeatedly developed and implemented, in response to the emergence of novel types of health technologies and the variability of the HTA context. However, with the development of HTA methods that focus on the use of RWD as an example, we have identified gaps in how HTA methods should be developed and implemented. Therefore, this thesis aimed to address these gaps, by conducting relevant conceptual research, and by illustrating how the conceptual research could help address the gaps, using RWD-related HTA methods as the cases.

This thesis was completed within the HTx project. HTx is a Horizon 2020 project supported by the European Union (65). One important objective of the HTx project is to facilitate the development of methodologies to deliver more customized information on the effectiveness and cost-effectiveness of complex and personalized combinations of health technologies. More specifically, the HTx project focuses on methodology for making real-world predictions of health outcomes at the population and individual level and methodology for personalized treatment. It thereby focuses on four disease areas: diabetes mellitus, head and neck cancer, multiple sclerosis and myelodysplastic syndrome (MDS). This thesis helps accomplish the HTx project objective, as our results could help HTA stakeholders within the project to better understand how to develop and implement the methods related to RWD. In other words, it could serve as a starting point for innovating the methods on RWD.

## Thesis outline

This thesis is divided in three parts. In the first part, we provide a conceptual framework to facilitate a general understanding of the process of innovating HTA methods and the stakeholder roles involved (Chapter 2). Also, we explore the applicability of this conceptual framework in three cases of innovating quantitative methods in various disease fields (such as diabetes and head and neck cancer), and improve the framework applicability by designing a roadmap in Chapter 3. In the second part (Chapters 4, 5 and 6) and the third part (Chapters 7 and 8), we investigate specific research questions related to development or implementation of qualitative methods and quantitative HTA methods using RWD, respectively. More specifically, in the second part, we focus

on methods used for assessing quality of studies using RWD, while in the third part, we focus on methods used for merging RCTs and RWD in (network) meta-analyses.

In chapter 4 (Part 2), a systematic review is conducted to assess methodological quality of retrospective observational studies investigating efficacy of diabetes monitoring systems. In this review, we apply the ROBINS-I tool, and investigate the methodologically quality change over time, by dividing the study into three subgroups according to publication year. In chapter 5 (Part 2), a literature review and content analysis is conducted to evaluate tools used to assess quality of real-world studies. In this study, we summarize signaling questions of all identified tools into quality items, using both deductive and inductive coding techniques, and score whether and to what extent a quality item is described by a tool. In chapter 6 (Part 2), we conduct a systematic review that evaluates methodological quality and applicability of risk prediction models to the HTA context. We apply the PROBAST (Prediction model Risk Of Bias Assessment Tool) (66) to assess RoB, and use findings from Betts et al. 2019 , which summarized recommendations and criticisms of HTA agencies on cardiovascular risk prediction models (67), to assess model applicability for the purpose of HTA.

Chapter 7 (Part 3) assesses how a method (i.e. the power prior) is used to merge data from RCTs and RWD in three parallel network meta-analyses. In this research, we estimate and compare effect sizes and rankings and test whether assumptions related to missing data, model type or weight of non-randomized studies impact the estimated efficacy. Chapter 8 (Part 3) further assesses the three methods (i.e., naïve pooling, power prior, and hierarchical modelling) to merge data from RCTs and RWD, by applying these methods with the "Crossnma" R package in four cases: two case studies on myelodysplastic syndromes and two on diabetes. After these three parts, we summarize the findings related to HTA method development and implementation, discuss the contribution of this thesis to HTA and the general healthcare system, and propose opportunities for future research, in the final "General Discussion" chapter.

## Author contribution

LJ wrote and edited the introduction. The supervisory team provided feedback throughout the process and approved the final version.

# References


1. O'Rourke B, Oortwijn W, Schuller T. The new definition of health technology assessment: A milestone in international collaboration. Int J Technol Assess Health Care. 2020 Jun;36(3):187-90.
2. Vokó Z, Cheung KL, Józwiak-Hagymásy J, et al. Similarities and differences between stakeholders' opinions on using Health Technology Assessment (HTA) information across five European countries: results from the EQUIPT survey. Health Res Policy Syst. 2016 Dec;14(1):1-7.
3. Jain B, Hiligsmann M, Mathew JL, Evers SM. Analysis of a small group of stakeholders regarding advancing health technology assessment in India. Value Health Regional Issues. 2014 May 1;3:167-71.
4. Williams AH, Cookson RA. Equity–efficiency trade-offs in health technology assessment. Int J Technol Assess Health Care. 2006 Jan;22(1):1-9.
5. Culyer AJ, Bombard Y. An equity framework for health technology assessments. Med Decis Making. 2012 May;32(3):428-41.
6. Haverinen J, Keränen N, Falkenbach P, Maijala A, Kolehmainen T, Reponen J. Digi-HTA: Health technology assessment framework for digital healthcare services. Finnish Journal of eHealth and eWelfare. 2019 Nov 2;11(4):326-41.
7. Bélisle-Pipon JC, Couture V, Roy MC, Ganache I, Goetghebeur M, Cohen IG. What makes artificial intelligence exceptional in health technology assessment?. Front Artif Intell. 2021 Nov 2;4:153.
8. Garrido MV, Gerhardus A, Røttingen JA, Busse R. Developing health technology assessment to address health care system needs. Health Policy (New York). 2010 Mar 1;94(3):196-202.
9. Stevens A, Milne R, Burls A. Health technology assessment: history and demand. J Public Health. 2003 Jun 1;25(2):98-101.
10. Banta D. The development of health technology assessment. Health policy. 2003 Feb 1;63(2):121-32.
11. European Network for Health Technology Assessment (EUnetHTA). HTA Core Model. Available from: https://www.eunethta.eu/hta-core-model. [Accessed Jul 6, 2023].
12. Lampe K, Mäkelä M, Garrido MV, et al. The HTA core model: a novel method for producing and reporting health technology assessments. Int J Technol Assess Health Care. 2009 Dec;25(S2):9-20.
13. Kristensen FB, Lampe K, Wild C, Cerbo M, Goettsch W, Becla L. The HTA Core Model®—10 years of developing an international framework to share multidimensional value assessment. Value Health. 2017 Feb 1;20(2):244-50.
14. National Health Care Institute (ZIN). Guideline for economic evaluations in healthcare. Available from: https://english.zorginstituutnederland.nl/publications/reports/2016/06/16/guideline-for-economic-evaluations-in-healthcare. [Accessed Nov 11, 2023]
15. National Institute for Health and Care Excellence (NICE). NICE health technology evaluations: the manual. Available from: https://www.nice.org.uk/process/pmg36/chapter/economic-evaluation. [Accessed Nov 11, 2023]
16. National Health Care Institute (ZIN). Guideline for building cost-effectiveness models in R. Available from: https://english.zorginstituutnederland.nl/publications/publications/2022/12/15/guideline-for-building-cost-effectiveness-models-in-r. [Accessed Jul 6, 2023].
17. National Health Care Institute (ZIN). Guidance for Outcomes Research. Available from: https://english.zorginstituutnederland.nl/publications/reports/2008/12/01/guidance-for-outcomes-research. [Accessed Jul 6, 2023].
18. National Institute for Health and Care Excellence (NICE). Reviewing our methods for health technology evaluation: consultation. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/chte-methods-consultation. [Accessed Jul 6, 2023].

19. Rosselli D, Quirland-Lazo C, Csanádi M, et al. HTA implementation in Latin American countries: comparison of current and preferred status. Value Health Reg Issues. 2017 Dec 1;14:20-7.

20. Shinkins B, Yang Y, Abel L, Fanshawe TR. Evidence synthesis to inform model-based cost-effectiveness evaluations of diagnostic tests: a methodological review of health technology assessments. BMC Med Res Methodol. 2017 Dec;17:1-0.

21. Chilcott J, Tappenden P, Rawdin A, et al. Avoiding and identifying errors in health technology assessment models: qualitative study and methodological review. Health Technol Assess. 2010 Jan 1;14(25):iii-v.

22. Foroutan N, Tarride JE, Xie F, Levine M. A methodological review of national and transnational pharmaceutical budget impact analysis guidelines for new drug submissions. Clinicoecon Outcomes Res. 2018 Nov 26:821-54.

23. Oliveira MD, Mataloto I, Kanavos P. Multi-criteria decision analysis for health technology assessment: addressing methodological challenges to improve the state of the art. Eur J Health Econ. 2019 Aug 1;20:891-918.

24. Polisena J, Garritty C, Kamel C, Stevens A, Abou-Setta AM. Rapid review programs to support health care and policy decision making: a descriptive analysis of processes and methods. Syst Rev. 2015 Dec;4(1):1-7.

25. Lehoux P, Williams-Jones B. Mapping the integration of social and ethical issues in health technology assessment. Int J Technol Assess Health Care. 2007 Jan;23(1):9-16.

26. Assasi N, Schwartz L, Tarride JE, Campbell K, Goeree R. Methodological guidance documents for evaluation of ethical considerations in health technology assessment: a systematic review. Expert Rev Pharmacoecon Outcomes Res. 2014 Apr 1;14(2):203-20.

27. Droste S, Dintsios CM, Gerber A. Information on ethical issues in health technology assessment: how and where to find them. Int J Technol Assess Health Care. 2010 Oct;26(4):441-9.

28. Lee A, Skött LS, Hansen HP. Organizational and patient-related assessments in HTAs: State of the art. Int J Technol Assess Health Care. 2009 Oct;25(4):530-6.

29. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. Health Technol Assess. 1999 May 10;3(5).

30. European Patients' Academy on Therapeutic Innovation (EUPATI). Health Technology Assessment process: Fundamentals. Available form : https://toolbox.eupati.eu/resources/health-technology-assessment-process-fundamentals. [Accessed Jul 7, 2023].

31. Cambridge Dictionary. Meaning of development in English. Available from: https://dictionary.cambridge.org/dictionary/english/development. [Accessed July 13, 2023].

32. Cambridge Dictionary. Meaning of implementation in English. Available from: https://dictionary.cambridge.org/dictionary/english/implementation. [Accessed July 13, 2023].

33. Rahimi K, Bennett D, Conrad N, et al. Risk prediction in patients with heart failure: a systematic review and analysis. JACC Heart Fail. 2014 Oct;2(5):440-6.

34. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat. 2012 Apr;132:365-77.

35. Silver SA, Shah PM, Chertow GM, Harel S, Wald R, Harel Z. Risk prediction models for contrast induced nephropathy: systematic review. Bmj. 2015 Aug 27;351.

36. van Giessen A, Peters J, Wilcher B, Hyde C, Moons C, de Wit A, Koffijberg E. Systematic review of health economic impact evaluations of risk prediction models: stop developing, start evaluating. Value Health. 2017 Apr 1;20(4):718-26.

37. Grant SW, Collins GS, Nashef SA. Statistical Primer: developing and validating a risk prediction model. Eur J Cardiothorac Surg. 2018 Aug 1;54(2):203-8.
38. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality.J Am Med Inform Assoc. 2021 Oct 1;28(10):2251-7.
39. Shenhar AJ, Holzmann V, Melamed B, Zhao Y. The challenge of innovation in highly complex projects: What can we learn from Boeing's Dreamliner experience?. J. Proj. Manag. 2016 Apr;47(2):62-78.
40. Shenhar AJ, Dvir D. Project management research—The challenge and opportunity. J. Proj. Manag. 2007 Jun;38(2):93-9.
41. Igelström E, Campbell M, Craig P, Katikireddi SV. Cochrane's risk of bias tool for non-randomized studies (ROBINS-I) is frequently misapplied: a methodological systematic review. J Clin Epidemiol. 2021 Dec 1;140:22-32.
42. D'Andrea E, Vinals L, Patorno E, et al. How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools. BMJ Open. 2021 Mar 1;11(3):e043961.
43. Quigley JM, Thompson JC, Halfpenny NJ, et al. Critical appraisal of nonrandomized studies—a review of recommended and commonly used tools. J Eval Clin Pract. 2019 Feb;25(1):44-52.
44. Schmitz S, Adams R, Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. Stat Med. 2013 Jul 30;32(17):2935-49.
45. Efthimiou O, Mavridis D, Debray TP, et al. Combining randomized and non-randomized evidence in network meta-analysis. Stat Med. 2017 Apr 15;36(8):1210-26.
46. Xie RZ, Malik ED, Linthicum MT, Bright JL. Putting stakeholder engagement at the center of health economic modeling for health technology assessment in the United States. Pharmacoeconomics. 2021 Jun;39(6):631-8.
47. Squires H, Chilcott J, Akehurst R, Burr J, Kelly MP. A framework for developing the structure of public health economic models. Value Health. 2016 Jul 1;19(5):588-601.
48. Breeze PR, Squires H, Ennis K, et al. Guidance on the use of complex systems models for economic evaluations of public health interventions. Health economics. 2023 Apr 20.
49. Branski RC, Cukier-Blaj S, Pusic A, et al. Measuring quality of life in dysphonic patients: a systematic review of content development in patient-reported outcomes measures. Journal of voice. 2010 Mar 1;24(2):193-8.
50. Goeree R, He J, O'Reilly D, et al. Transferability of health technology assessments and economic evaluations: a systematic review of approaches for assessment and application. Clinicoecon Outcomes Res. 2011 Jun 2:89-104.
51. Németh B, Goettsch W, Kristensen FB, et al. The transferability of health technology assessment: the European perspective with focus on central and Eastern European countries. Expert Rev Pharmacoecon Outcomes Res. 2020 Jul 3;20(4):321-30.
52. Makady A, van Veelen A, Jonsson P, et al. Using real-world data in health technology assessment (HTA) practice: a comparative study of five HTA agencies. Pharmacoeconomics. 2018 Mar;36:359-68.
53. Makady A, Ten Ham R, de Boer A, Hillege H, Klungel O, Goettsch W. Policies for use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies. Value Health. 2017 Apr 1;20(4):520-32.
54. Griffiths EA, Macaulay R, Vadlamudi NK, Uddin J, Samuels ER. The role of noncomparative evidence in health technology assessment decisions. Value Health. 2017 Dec 1;20(10):1245-51.
55. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W. What is real-world data? A review of definitions based on literature and stakeholder interviews. Value health. 2017 Jul 1;20(7):858-65.
56. Hall PS. Real-world data for efficient health technology assessment. Eur J Cancer. 2017 Jul 1;79:235-7.

57. Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling good research practices—overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–1. Med Decis Making. 2012 Sep;32(5):667-77.

58. Regier DA, Pollard S, McPhail M, et al. A perspective on life-cycle health technology assessment and real-world evidence for precision oncology in Canada. NPJ Precis Oncol. 2022 Oct 25;6(1):76.

59. Schünemann HJ, Tugwell P, Reeves BC, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. Res Synth Methods. 2013 Mar;4(1):49-62.

60. Jonathan AC Sterne, Miguel A Hernán, Alexandra McAleenan, Barnaby C Reeves, Julian PT Higgins. Assessing risk of bias in a non-randomized study. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Cochrane;2022 Chapter 25. Available from: https://training.cochrane.org/handbook/current/chapter-25. [Accessed Feb 8 2023].

61. Griffiths EA, Vadlamudi NK. Not ready for the real world? The role of non-RCT evidence in health technology assessment. Value Health. 2016 May 1;19(3):A286.

62. Pease A, Lo C, Earnest A, Kiriakova V, Liew D, Zoungas S. The efficacy of technology in type 1 diabetes: a systematic review, network meta-analysis, and narrative synthesis. Diabetes Technol Ther 2020 May 1;22(5):411-21.

63. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016 Oct 12;355.

64. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. Stat Med. 2015 Dec 10;34(28):3724-49.

65. HTx. About HTx project. Available from: https://www.htx-h2020.eu/about-htx-project. [Accessed Oct 25, 2022].

66. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S, PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019 Jan 1;170(1):51-8.

67. Betts MB, Milev S, Hoog M, et al. Comparison of recommendations and use of cardiovascular risk equations by health technology assessment agencies and clinical guidelines. Value Health. 2019 Feb 1;22(2):210-9.

1

# Part 1

## Conceptual Framework for HTA Methods Innovation

Chapter 2

# Understanding Innovation of Health Technology Assessment Methods – the IHTAM Framework

Li Jiu, Milou A Hogervorst, Rick A Vreman, Aukje K Mantel-Teeuwisse, Wim G Goettsch

*International Journal of Technology Assessment in Health Care. 2022;38(1):e16.*

# Abstract

## Background

Adequate methods are urgently needed to guarantee good practice of health technology assessment (HTA) for technologies with novel properties. The aim of the study was to construct a conceptual framework to help understand Innovation of HTA Methods (IHTAM).

## Methods

The construction of the IHTAM framework was based on two scoping reviews, one on current practice of innovating methods, i.e. existing HTA frameworks, and one on theoretical foundations for innovating methods outside the HTA discipline. Both aimed to identify and synthesize concepts of innovation (i.e. innovation processes and roles of stakeholders in innovation). Using these concepts, the framework was developed in iterative brainstorming sessions and subsequent discussions with representatives from various stakeholder groups.

## Results

The framework was constructed based on twenty documents on innovating HTA frameworks and fourteen guidelines from three scientific disciplines. It includes a generic innovation process consisting of three phases ("Identification", "Development", "Implementation") and nine subphases. In the framework, three roles that HTA stakeholders can play in innovation ("Developers", "Practitioners", "Beneficiaries") are defined and a process on how the stakeholders innovate HTA methods is included.

## Conclusions

The Innovation of HTA Methods framework visualizes systematically which elements and stakeholders are important to the development and implementation of novel HTA methods. The framework could be used by all stakeholders involved in HTA innovation to learn how to engage dynamically and collaborate effectively throughout the innovation process. HTA stakeholders in practice have welcomed the framework, though, additional testing its applicability and acceptance is essential.

## Introduction

Health technology assessment (HTA) has become increasingly important throughout the world as a process to systematically evaluate properties and effects of a health technology with the purpose of supporting evidence-based decision making in reimbursement and clinical treatment (1). To guarantee good practices in HTA, adequate HTA methods are needed (2). HTA methods refer to all qualitative and quantitative methods relevant to the full scope of the HTA process (3), such as methods for evidence generation from clinical or real-world data (4,5), methods for synthesizing HTA evidence and modelling cost-effectiveness (6), and tools for dealing with uncertainty in multi-criteria decision-making for healthcare (7). These methods vary by function and, if proven robust and implemented successfully, can improve the quality of HTA conducted throughout the HTA process.

The need for novel HTA methods becomes urgent when existing methods are not able to handle complexity of emerging health technologies, which creates barriers for a systematic evaluation. Novelty here refers to the quality of being unusual in either structure or content of an HTA method, with the potential to resolve conflicts between traditional methods that HTA relies on and quality of the HTA for emerging health technologies (8,9). For example, genetic testing, an emerging health technology to prognose individuals with high risks of genetic diseases, is ethically complex, so novel methods are needed to measure and value its ethical issues in HTA decision-making (10). Digital health, another example of new technologies with unique features in data security and artificial intelligence, also needs specially designed methods to define and evaluate its HTA-related evidence (11).

To satisfy the urgent needs, HTA methods are developed and implemented, in other words, innovated, mainly in two ways: creation based on multiple disciplines of knowledge and improvement based on previously innovated methods. As the number of innovated methods increases dramatically, guidelines, such as the HTA core model (12), have been applied to inform HTA stakeholders (e.g. academics, healthcare professionals, HTA bodies, governments, patients, payers, and industry) on how to select HTA methods for different technologies in different settings. However, HTA stakeholders still lack an understanding of how to create or improve HTA methods. Consequently, stakeholders, especially those without an HTA knowledge background (such as patients and healthcare professionals), may lack consensus on which methods are urgently needed, how to innovate them, and, equally importantly, how they could engage in the innovation.

Therefore, the objective of this study was to develop a framework with two functions: to illustrate a generic innovation process that is applicable to all types of HTA methods; and to illustrate how different HTA stakeholder groups can engage dynamically and collaborate effectively throughout the innovation process. We adopted a conceptual framework approach, which defines a network of concepts providing comprehensive understanding of multidisciplinary phenomena and helps stakeholders understand knowledge from other disciplines (13). We considered this approach most useful to facilitate understanding of the complexities associated with innovating HTA methods.

## Methods

The new framework was developed in two stages: first, identifying and synthesizing concepts of innovating HTA methods in two scoping reviews; and second, drafting the framework based on the concepts and refining the framework by gaining input from HTA stakeholders in the HTx project. This is an ongoing research project funded under the Horizon 2020 Framework Programme, with the aim to support patient-centered, societally oriented, real-time decision-making for integrated healthcare throughout Europe (14). The flow diagram of developing the Innovation of HTA Methods (IHTAM) framework can be found in Figure 1.

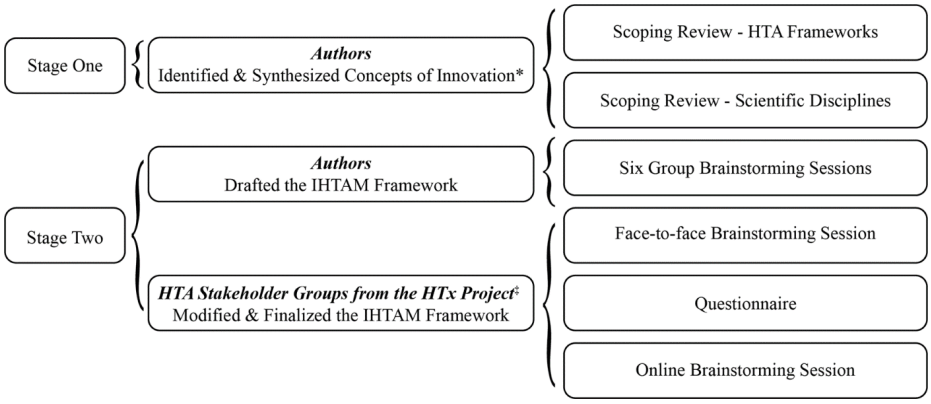



**Figure 1. Flow diagram of constructing the IHTAM framework.**

* Concepts of innovation indicate innovation processes and roles of stakeholders in innovation;

‡ A research project with an aim to develop and implement novel methods for patient-centered decision-making using real-world data and machine learning techniques.

## Identifying and Synthesizing Concepts of Innovation (stage one)

Our starting point was to identify concepts of innovation, defined as processes of innovation and stakeholders involved. Such concepts were considered likely to occur in two sources, therefore we performed two scoping reviews. The first source was literature on innovating HTA methods. Since we expected that lots of methods were innovated in the past through a variety of formats (e.g. frameworks, models, tools) and that the concepts extracted from different formats shared similarities, we limited ourselves to reviewing HTA frameworks. The second source was literature from scientific disciplines (defined as branches of knowledge) relevant to innovation, which might provide theoretical foundations for innovating HTA methods. For the two scoping reviews, we drafted protocols following PRISMA guidance (15) and conducted a pilot test to refine eligibility criteria, search strategies, and processes of data screening, abstraction, and synthesis.

### *Scoping review on HTA frameworks*

HTA frameworks were identified in both scientific articles and grey literature. Documents were searched from PubMed, Embase, and Google Scholar. The search strategy included "framework" and "health technology assessment" (or "HTA") in title and/or abstract. An article was included if it described a process in the methodology part on how an HTA method was developed, implemented, validated, or transferred; and excluded if it was not in English or full text was not available. The complete search strategy appears in Appendix 1. According to the same in- and exclusion criteria, grey literature was searched from Google Advanced Search and websites of seven international organizations which might report innovation of HTA methods, including the World Health Organization (WHO), the European Network for Health Technology Assessment (EUnetHTA), the Professional Society for Health Economics and Outcomes Research (ISPOR), the Society for Health Technology Assessment International (HTAi), the International Network of Agencies for Health Technology Assessment (INAHTA), the Institute for Clinical and Economic Review (ICER), and the National Institute for Health and Care Excellence (NICE). We searched for "HTA framework" and took the first twenty items (sorted by relevance) of grey literature from each source because a pilot test showed that the first ten items were most likely to be eligible. Citations in eligible scientific articles and grey literature were also scanned for eligibility. Data screening was independently conducted by one author (LJ) and cross-checked (10%) by another (MH).

Data items extracted from eligible studies included study characteristics (i.e. first author, publication year), a description of the innovation processes, and the stakeholders involved with their roles. Subsequently, data items regarding innovation

processes or stakeholders involved were clustered and items with similar meanings were merged. For example, a process of "prototyping methods" and a process of "drafting solutions to a problem" were clustered as "design prototypes"; doctors and nurses were all clustered as healthcare professionals. Data items were extracted and clustered by one reviewer (LJ) and a random subset (ten percent) was checked by another (MH). Any discrepancies in data screening or extraction were resolved by discussion.

***Scoping review of scientific disciplines on innovation***

Concepts from scientific disciplines on innovation were identified only in scientific articles because a pilot search failed to identify eligible results in grey literature. An article was included if it provided a guideline on how to develop, implement, validate or transfer an object; and excluded if the guideline was tailored to a specific object (e.g. school psychology), not in English, or full text was not available. The strategy of searching for concepts within scientific disciplines on innovation was also identical to that of the previous review, except for the search terms used (i.e. innovation, identification, research, development, implementation, validation, transfer, generalization). Given the large number of items listed by databases, we only scanned the first 200 items (sorted by relevance) of each database as the pilot test showed data items after fifty of each database became less relevant. After identifying eligible articles, we further clustered them based on scientific disciplines. By scanning titles and abstracts of each article, we could identify theoretical foundations of the innovation processes, and then determined which discipline an article belongs to. For example, a "framework for design thinking in health innovation" and a "design thinking framework for healthcare management and innovation" were clustered into a discipline called "design thinking". The processes of data screening, abstraction, and clustering were also identical to those of in the first review.

## Drafting and Refining the IHTAM Framework (stage two)

***Brainstorming Sessions***

Based on results of the reviews, the five authors organized six brainstorming sessions in three consecutive weeks to construct the framework. All opinions were recorded into notes by LJ and reconfirmed by the authors who expressed them. Axial coding was used to identify how the concepts regarding innovation processes and those regarding stakeholder roles interact with each other, in other words, what roles HTA stakeholders could play and how their roles change along different phases of innovation. Selective coding was then used to select overarching concepts which all authors agreed to capture the essence of innovating HTA methods. Concepts without enough supporting data were deleted.

***Stakeholder input from the HTx Project***

To further refine the draft framework, two further sessions, one face-to-face and one online, were organized on 7th February, 2020 and 30th June, 2020, respectively, during the HTx consortium meetings. All the participants of the consortium meeting received a notification of the rationale and schedule of the sessions one week before and were asked to confirm participation. The attendants were presented the latest version of the draft framework and asked to judge relevance of the conceptualized innovation phases and stakeholder roles to the real-world practice of innovating HTA methods. Before the session, a questionnaire with open questions was sent to the attendants for preparation and clarification of their opinions. LJ recorded all the attendants' opinions into notes and sent them e-mails for reconfirmation in case of any uncertainty. Open, axial, and selective coding was applied by LJ to conceptualize the notes. To avoid the subjective coding bias, the coding process was reviewed by RV.

# Results

## Identifying and Synthesizing Concepts of Innovating HTA Methods (stage one)

The flow diagram of identifying eligible studies and study characteristics of the two scoping reviews appears in Appendix 2 and 3. Phases of innovation and stakeholders involved in innovation from the two scoping reviews are shown in Table 1.

***Review on HTA Frameworks***

Twenty eligible documents (see Appendix 3) on innovating HTA frameworks were identified. The processes of innovation were clustered into nine phases (from "Identify needs for innovation" to "Transfer innovation"), and HTA stakeholders involved in innovation were clustered into seven categories: academics (mentioned most frequently, in 95% of the documents), healthcare professionals, HTA bodies, governments, patients, payers, and industry. In each phase of innovation, various categories of HTA stakeholders were involved, but we did not identify a pattern in distribution of different HTA stakeholders across these phases. For example, of the five documents (25% of all identified) mentioning patient groups being involved in innovation, one disseminated a method (16); three tested HTA methods in case studies (17-19); and one evaluated method performance in practice (20).

**Table 1. Phases of innovation and stakeholders involved in innovation from the two scoping reviews**

| | Identify needs for innovation | Collect resources needed for innovation | Prototype innovation | Test innovation in case studies | Disseminate innovation | Make decisions to adopt innovation | Implement innovation | Evaluate innovation in practice | Transfer innovation | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Documents on innovating HTA frameworks (n=20)*** | | | | | | | | | | |
| Academics | 9 | 16 | 18 | 9 | 4 | 0 | 4 | 3 | 4 | 19 (95%) |
| Healthcare professionals | 3 | 1 | 3 | 6 | 0 | 0 | 5 | 3 | 3 | 14 (70%) |
| HTA bodies | 4 | 1 | 3 | 1 | 3 | 0 | 2 | 1 | 1 | 8 (40%) |
| Governments | 1 | 2 | 2 | 2 | 2 | 0 | 2 | 1 | 2 | 5 (25%) |
| Patients | 2 | 0 | 3 | 2 | 1 | 0 | 1 | 2 | 0 | 5 (25%) |
| Payers | 0 | 0 | 1 | 2 | 1 | 0 | 1 | 0 | 0 | 3 (15%) |
| Industry | 1 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 2 (10%) |
| Total | 10 | 16 | 18 | 13 | 7 | 0 | 6 | 3 | 4 | 20 (100%) |
| ***Documents on scientific disciplines on innovation (n=14)*** | | | | | | | | | | |
| Developers | 9 | 10 | 8 | 1 | 5 | 0 | 6 | 5 | 2 | 11 (79%) |
| Practitioners | 7 | 10 | 0 | 2 | 5 | 0 | 7 | 5 | 2 | 10 (71%) |
| Community [a] | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 7 (50%) |
| Decision makers [b] | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 (14%) |
| Planners [c] | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 4 (28%) |

**Table 1. Continued**

| | Identify needs for innovation | Collect resources needed for innovation | Prototype innovation | Test innovation in case studies | Disseminate innovation | Make decisions to adopt innovation | Implement innovation | Evaluate innovation in practice | Transfer innovation | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Documents on scientific disciplines on innovation (n=14)*** | | | | | | | | | | |
| Technical assistance experts [d] | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 (7%) |
| Policy makers [e] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 (7%) |
| Total | 9 | 10 | 8 | 2 | 6 | 2 | 8 | 5 | 2 | 14 (100%) |

[a] Community indicate stakeholders who have problems and need innovative solutions;
[b] Decision-makers, who decide on whether to adopt innovation;
[c] Planners, who consider contexts and stakeholders responsible for program adoption, implementation, and adoption;
[d] Technical assistance experts, who record implementation progress and advise on how to improve implementation processes;
[e] Policy makers, who develop policies regarding innovation.

***Review on Scientific Disciplines on Innovation***

Fourteen eligible documents from three scientific disciplines on innovation (design thinking (n=4), implementation research (n=9), and interdisciplinary research (n=1)) were identified (Appendix 3). Innovation processes identified in this body of literature could be clustered into nine phases. Eight of the nine were similar to those from HTA frameworks, except for making decisions to adopt innovation (21-23), which was not mentioned by any HTA framework. In addition, compared to HTA frameworks, the three disciplines outside HTA provided more clarity on implications of each innovation phase. For example, innovation guidance from the discipline of design thinking implied that developers may observe other stakeholders' behavior when identifying needs (24); guidance from implementation research implied that innovation should be disseminated clearly and concisely to stakeholders in various user-friendly formats (25). These detailed implications were not mentioned in HTA frameworks.

In the three disciplines, stakeholders were clustered into seven categories based on their roles in innovation. The most mentioned categories (developers, practitioners, and community) occurred across phases of innovation while the less mentioned categories (decision-makers, planners, technical assistance experts, and policy makers) occurred only in the last five phases (from "Disseminate innovation" to "Transfer innovation").

## Drafting the IHTAM Framework and Refining the Framework by gaining input from stakeholders (HTx project) (stage two)

Seven HTA stakeholders attended the face-to-face brainstorming session and six attended the online session. One stakeholder did not attend the sessions but completed the questionnaire for the online session. In all the fourteen stakeholders, academia accounted for eight, while representatives of HTA bodies, representatives of industry, and patients each accounted for two.

***Phases of Innovation***

All meeting participants considered the innovation phases from the two reviews relevant to innovating HTA methods, but some phases could be further split (e.g. "Identify needs for innovation") or merged (e.g. "Disseminate innovation"), to be more understandable for them. They also advised to cluster the framework into three main phases, called "Identification", "Development", and "Implementation", and to explain what tasks should be resolved through defining multiple subphases within each phase.

***Roles of Stakeholders***

The classic way of describing HTA stakeholders, e.g. HTA bodies, payers, patients, and industry, does not specify the roles they may take within innovation processes. In

contrast, the categories of stakeholders that were derived from the scoping review of scientific disciplines are more widely applicable and better fit the purpose of roles of stakeholders within general guidance for innovation. One classic HTA stakeholder may take different roles within an innovation process. For example, academics could act as developers in one phase and as practitioners in another. Healthcare professionals could act as practitioners but also as decision-makers.

To retrieve a small set of generic stakeholder roles for the innovation process, we further clustered the roles from the two reviews. Decision-makers and technical assistance experts were considered being developers or practitioners; policy makers and community were not directly involved in developing or implementing innovation, but were affected by innovation, so we clustered them into "beneficiaries". We thus defined beneficiaries, developers, and practitioners as the three generic roles that HTA stakeholders could play in innovating HTA methods, as shown in Table 2.

**Table 2. Definitions of generic stakeholder roles in innovation**

| Generic roles | Definitions | Examples |
|---|---|---|
| Beneficiaries | Stakeholders who benefit from or are affected by HTA methods | HTA bodies, healthcare professionals, patients, and industry who proposed limitations of existing methods and recommended innovation of novel methods (16) |
| Developers | Stakeholders who develop HTA methods | Academics who analyzed feedback from other stakeholders and revised a method (26;27) |
| Practitioners | Stakeholders who implement and use HTA methods | Healthcare professionals or policy makers who evaluated how to tailor a method to local contexts and whether the tailored method could be adopted (28;29) |

The HTx meeting participants agreed in principle that developers, practitioners, and beneficiaries could be tailored to contexts where HTA methods were innovated. But they emphasized that, in addition to the final framework illustrating an innovation process and stakeholders roles, it needed to be made explicit how, in general, the classic categories of stakeholders, such as HTA bodies and patients, would translate to the stakeholder roles. After coding from the meeting participants' opinions, we defined how HTA stakeholders engage in innovation, as shown in Figure 2. HTA stakeholders can do so through two phases, which are called "role recognition" and "stakeholder discovery".

Role recognition indicates that HTA stakeholders first need to realize their roles in each phase of HTA method innovation. Stakeholder discovery indicates that, for each subphase, HTA stakeholders already involved in innovation may discover additional HTA stakeholders who are qualified as beneficiaries, practitioners, or developers. The stakeholders may, based on their own experience, evaluate who may be qualified for the three roles. After evaluation, those potentially qualified may be invited and contribute to the innovation. Since the tasks of beneficiaries, practitioners, and developers vary in different subphases of innovation, "role definition" and "stakeholder discovery" should be conducted iteratively throughout the innovation process.

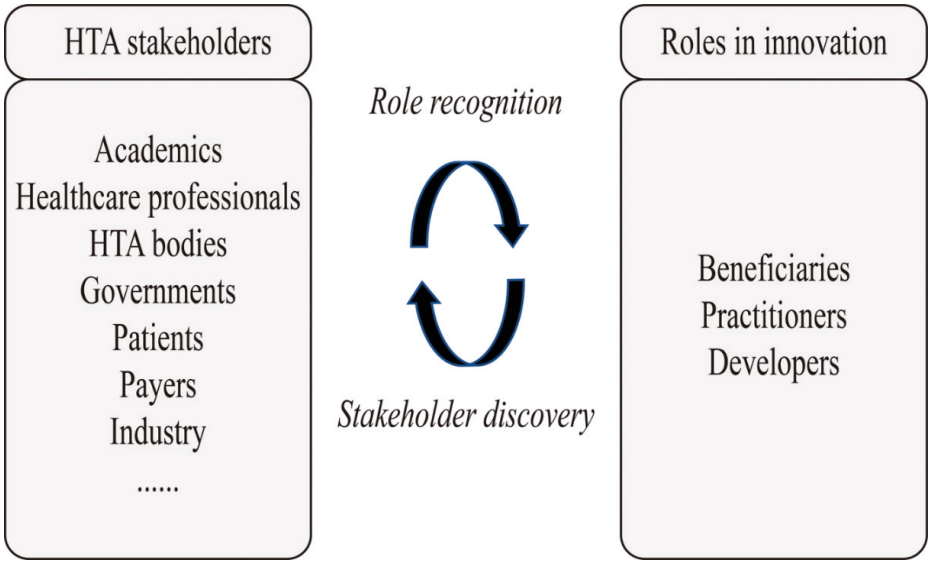



**Figure 2. A process on how HTA stakeholders can engage in innovating HTA methods.**

This figure illustrates how to engage HTA stakeholders in innovating HTA methods. The box on the left indicates HTA stakeholders (e.g. academics, HTA bodies) that can engage in innovation. The ellipsis at bottom left indicates engagement of additional HTA stakeholder groups is also possible. The box on the right indicates the three roles HTA stakeholders can play in innovation ("Beneficiaries", "Practitioners", "Developers"). In the middle of the concept map lists a two-phase process ("role recognition" and "stakeholder discovery") on how HTA stakeholders play the three roles of innovation.

## The Final Framework

The final framework is shown in Figure 3 and illustrates a generic innovation process of HTA methods with three phases (i.e. "Identification", "Development", "Implementation"). The three phases are distinguished by three colors, and each phase includes three subphases in white boxes. Underneath each subphase the roles HTA stakeholders can play in that subphase are noted.

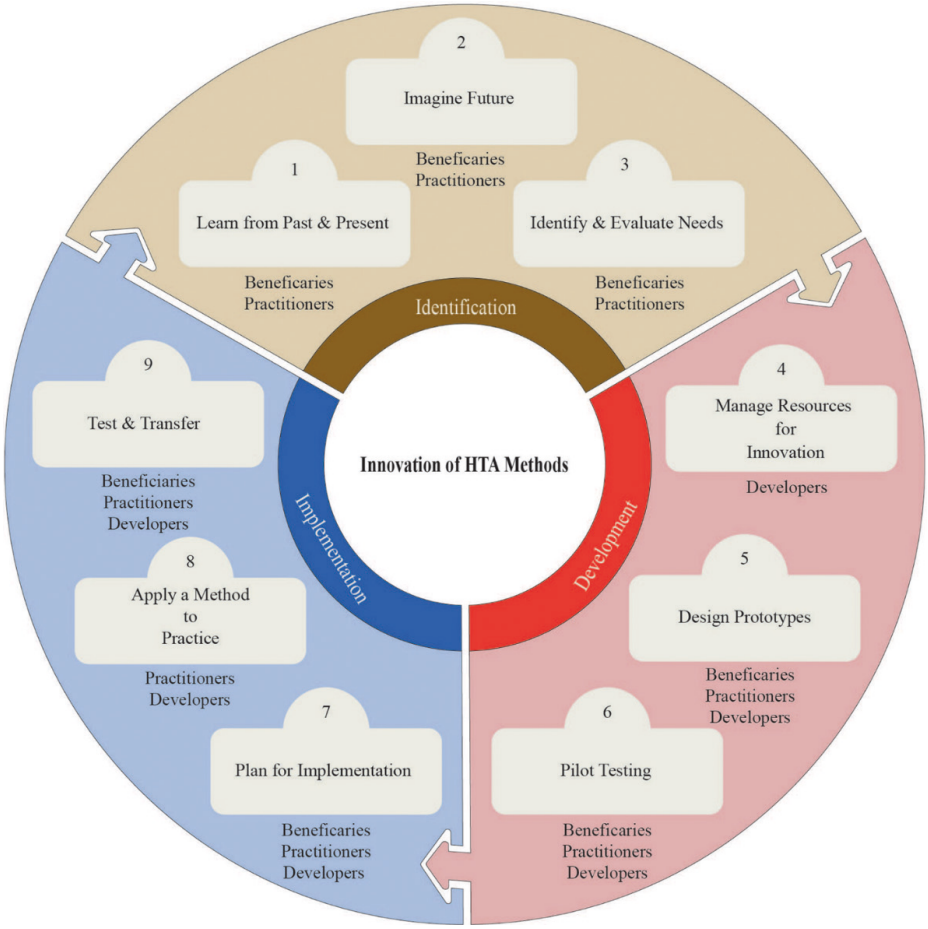



**Figure 3. A generic process on how HTA methods are innovated.**

This concept map illustrates all key concepts of the IHTAM framework: (1) concepts relevant to a generic innovation process with three phases (i.e. "Identification", "Development", "Implementation", which are distinguished by three colors) and nine subphases (in white boxes with numbers from 1 to 9); (2) roles of HTA stakeholders in innovation in each subphase (attached under each white box).

***Phase One - Identification***

The identification phase, as the first phase of innovation, rationalizes the HTA method innovation and justifies stakeholders to be involved. In this phase, HTA stakeholders learn from past and present, imagine the future, and identify and evaluate the needs. "Learning from past and present" indicates that the stakeholders should acquire insight in limitations of current HTA processes. The commonly used techniques include surveys, interviews, literature reviews, or observations on how an HTA progress is conducted (11,30 - 32). A recommendation on how to identify up-to-date limitations is to gain feedback from practitioners who used traditional methods and beneficiaries who are affected by them. As emphasized by the design thinking theory, stakeholders may not really realize a limitation themselves (33,34). Still, limitations may be identified after observing and analyzing how practitioners act in practice (35). "Imagine Future" refers to picturing what future HTA processes looks like, and identifying enablers and barriers for the imagined future. One way to achieve this is to construct future scenarios through round-by-round brainstorming with the techniques such as group interviews and surveys (11,31,36). Future imagining could be conducted together with learning from past and present. Identification and evaluation of needs, as the third subphase, is the goal of the identification phase and the premise of developing HTA methods. Based on a gap identified by comparing future HTA scenarios with current HTA practices, HTA stakeholders may evaluate heterogeneity of contexts where gaps are identified. The various contexts in which HTA is conducted, such as different types of health technologies, disease areas, or geographic areas, need to be considered as corresponding needs may vary. Once needs are identified, stakeholders may decide whether existing methods can be improved or novel methods need to be developed. A decision could be made by investigating transferability opportunities, as suitable methods may already exist in other contexts. The methods innovated originally in other disciplines of knowledge may be worth studying if they have potential to be applied in HTA. A challenging task throughout the identification phase is the participation from a large group of stakeholders with different roles. Not only academics, but also any potential stakeholders qualified as potential practitioners and beneficiaries could identify or evaluate the needs. In practice, stakeholders except academics are less involved in needs identification or evaluation (see Table 1). Our suggestion is adopting a regular procedure of "stakeholder discovery" and "role definition", as illustrated in Figure 3. In this way, initial involved stakeholders, e.g. academics, could identify and invite other stakeholders with clarified distinguished roles.

***Phase Two - Development***

To develop an HTA method robustly, several concerns should be considered. First, resources for innovation should be managed in a good way. This usually begins with

human resource management, that is, defining a group of method developers from a range of HTA stakeholders. Developers may set the priority for the needs that a novel method addresses, then establish an external research communication mechanism and avoid duplication of efforts of development. Academics could lead the group of developers, but other HTA stakeholders could also take up the role, depending on the contexts (18,37). Then developers should make agreements on the concentration and allocation of all the other resources, such as time, finance, and knowledge (23,25,38). A typical way of resource management is to conduct a feasibility analysis to evaluate what resources are needed and whether resources are available (39). Second, if method development is feasible, developers may design a method prototype and its derivative versions based on heterogeneity of needs, to improve the method capability that can be transferred to various HTA contexts. Feedback from practitioners and beneficiaries should also be reflected in method development, as innovation successes largely depend on how easily a method can be implemented (33). Therefore, developers need systematic approaches of gaining feedback regularly from beneficiaries and practitioners. One solution could be "ideation", a commonly used process in the design thinking theory, which synthesizes insights from multiple stakeholders for addressing design challenges (40,41). The final subphase of development "Pilot testing" is to validate HTA method prototypes. Before applying the prototypes to practice, developers may first disseminate method prototypes to practitioners and engage those who feel interested in the methods being developed. These practitioners then implement methods in pilot contexts (16,36,37). One concern is how to identify and organize pilot case studies which could simulate real-world practice while avoiding consequences in case of any error caused by design flaws, lack of transferability, or wrong operations. Method validity could be judged by all stakeholders in a structural way (27,29,32,37).

***Phase Three - Implementation***

A method innovation process is not complete until a method is implemented successfully. During the implementation, as what implementation science often stresses, stakeholders need to plan for implementation, apply a method to practice, then transfer it to other contexts after validation (34). Any developer or practitioner involved in method development may contribute to diffusion (e.g. scientific publications and conferences) or dissemination (e.g. training) of methods to practitioners in real-world practice (34). Implementation strategies may also be developed, in which all resources needed for conducting and monitoring implementation are considered (24,38,41). Strategies need be tailored for different contexts where HTA is conducted. One challenge of planning for implementation is how to motivate real-world practitioners

and beneficiaries to adopt the novel method in practice, as any reluctance to method uncertainty or misunderstanding could deter the adoption.

Once a method is adopted, concerted effort is required by all stakeholders who are qualified as practitioners to implementing the method (37). Developers, with knowledge of a novel method, should continuously provide technical assistance and work with practitioners to adjust implement strategies to various contexts when necessary(22,42,43). A feedback loop, which cycles through the method application by monitoring, adoption, and tailoring, could make an HTA method more sustainably entrenched within a context (37). Regular debriefing of implementation progresses could be performed for the later validation purpose (23).

Finally, in the last subphase "Test & Transfer", performance of a method should be test with an intention of further innovation. Developers need sound approaches to systematically test the validity of HTA methods, then report the results transparently to all stakeholders. The results worth reporting include outcomes of an HTA method, the extent to which a method is adopted by practitioners and beneficiaries, and quality of implementation strategies (21-23). Practitioners from other contexts may be invited, as they could help judge method transferability and point out potential concerns during the transfer. Group decision-making is required on whether the method is robust, and in what condition it can be transferred (22). Finally, discussion may be initiated to justify the necessity for another round of innovation.

## Discussion

We developed a conceptual framework which provides an understanding of how to innovate HTA methods. The IHTAM framework illustrates a generic innovation process on how to identify needs for, develop, and implement HTA methods. The framework also outlines a process on how HTA stakeholders can engage in innovating HTA methods.

Our framework adds value to HTA good practice for several reasons. First, the framework contributes to collaboration of HTA stakeholders from various disciplines. By defining three generic roles (beneficiaries, practitioners, developers) of innovation and tasks of each role in each phase of innovation, the framework prompts HTA stakeholders to think beyond the traditional view on stakeholder roles whether, at which phase(s), and for which role(s) they are qualified for innovation.

Second, as the first to provide a general understanding of innovating HTA methods, the framework serves a foundation for constructing or improving more specific guidance on innovation. Some specific guidance does already exist. For example, a guideline was developed for developing, implementing, evaluating, and reporting discrete event simulation (DES), a novel computer-based modelling that is increasingly applied in the HTA context (44). The guideline described relevant concerns and best practice recommendations throughout the innovation process. Another example is a report developed by ISPOR to guide developing and implementing a type of HTA decision-making method - multiple criteria decision analysis (MCDA) - to support healthcare decisions (45). The report outlines an eight-phase process of MCDA development and implementation. While these guides focus on innovation of one type of HTA method, our framework provides a general understanding of innovating all types of methods.

Third, the framework promotes consideration of key challenges that may exist in innovating HTA methods. It lists phases of innovation which may be implicitly known but not explicitly considered currently by HTA stakeholders. For example, the subphase "Apply a Method to Practice" implies that practitioners may decide on whether to apply a method to practice. Apart from considering who are qualified as practitioners, HTA stakeholders in a specific context may consider what criteria should be used for decision-making. Attaching importance to challenges of innovation contributes to method validity and implementation success.

### How to use the IHTAM framework

The IHTAM framework has the potential to become a starting point for HTA stakeholders to understand their roles in innovating HTA methods and we consider all HTA stakeholders as potential audience of the framework. It is important to realize that (sub)phases of the IHTAM framework do not necessarily occur sequential and a specific innovation process should always be defined for each method innovated. To determine the most appropriate process and roles, stakeholders always need to consider actual conditions and initiate detailed discussion. The function of the IHTAM framework in determining an appropriate innovation process or roles is to explicitly illustrate what aspects of innovation need to be considered. In summary, we recommend considering the following when using the IHTAM framework:

1. Consider all (sub)phases and three innovation roles within the IHTAM framework and judge their relevance to the methods to be innovated;
2. Discuss whether additional (sub)phases or roles of innovation apply;
3. Construct a tailored innovation framework and consider challenges of innovation to be addressed;
4. Evaluate qualification of HTA stakeholders for innovation and facilitate collaboration.

## Limitations

One limitation of our study is that the article selection of grey literature may be difficult to replicate. In our search strategy, only the first twenty items of "HTA frameworks" listed on Google Advanced Search and the seven international organizations were included. The sequences of items from the above-mentioned sources can be influenced by their searching algorithms. Particularly, the Google Advanced Search is highly influenced by a user's own preference. However, this limitation would not cause much impact on the coding, as only one of the 34 eligible literature documents was sourced from the grey literature (Appendix 3). For data extraction of the scoping reviews, one reviewer independently scanned all titles and abstracts while the second reviewer checked only 10% of them. This might cause exclusion of some eligible literature but might not influence the results of conceptualization. Our reason is that the included literature, which repeatedly described the similar phases of innovation and stakeholder categories involved (Table 1), were already sufficient (n=34) for the coding purpose. Furthermore, there are several limitations on the framework applicability. For the review of HTA methods, only the "framework" type of methods was included, so the applicability of the IHTAM framework might be limited when applying to other types of methods, such as HTA models. For the brainstorming sessions, we did not invite HTA stakeholders outside the HTx project. Even within the project, we relied on a relatively low number of HTA stakeholders to confirm usefulness of the framework. Another limitation is that not all HTA stakeholder groups, such as payers, were invited for input. Hence, uncertainty still exists on whether the framework is accepted by HTA stakeholders in various contexts. Still, recommendations provided by the IHTAM framework are worth considering, because it can serve as a starting point to illustrate the complex innovation process and how it is related to HTA stakeholders. Although the IHTAM framework will not function as a quality checklist that can be rigidly followed, the way we conceptualize the method innovation and the relevant challenges we propose are worth noting for all types of HTA methods and for all HTA stakeholders.

We recommend to create an in-depth pathway based on the elements described in our framework for further identifying and solving particular challenges in innovation, which may ultimately contribute to a quality checklist. We also recommend future efforts to testing the applicability and acceptance of the IHTAM framework in case studies of innovating HTA methods in various contexts.

## Conclusions

The IHTAM framework provides an understanding of how to innovate HTA methods and it helps HTA stakeholders better understand how to engage in innovation by knowing what different roles they can play in complex contexts of innovation. We believe the framework may add value to development of robust HTA methods and effective implementation, which helps meet the needs for novel HTA methods due to emerging health technologies.

## Author contribution

LJ conducted the scoping reviews, developed the first draft of the conceptual framework, obtained feedback on the framework from the other authors and stakeholders in the HTx project, and wrote the draft manuscript. All co-authors contributed to this study and critically reviewed and approved the manuscript.

# References


1. Health Technology Assessment International (HTAi). Understanding Health Technology Assessment (HTA) Available from: https://htai.org/wp-content/uploads/2018/02/PCISG-Resource HEE_ENGLISH_PatientGuidetoHTA_Jun14.pdf. [Accessed Feb 26, 2020].
2. Kristensen FB, Husereau D, Huić M, et al. Identifying the need for good practices in health technology assessment: Summary of the ISPOR HTA Council Working Group Report on Good Practices in HTA. Value Health. 2019;22(1):13-20.
3. World Health Organization (WHO). Health technology assessment glossaries. Available from: https://www.who.int/health-technology-assessment/about/Glossaries/en. [Accessed Jan 19, 2021].
4. Curtis JR, Foster, PJ, Saag KG. Tools and Methods for Real-World Evidence Generation: Pragmatic Trials, Electronic Consent, and Data Linkages. Rheumatic Disease Clinics. 2019; 45(2): 275-289.
5. Ridyard CH, Hughes DA. Methods for the collection of resource use data within clinical trials: a systematic review of studies funded by the UK Health Technology Assessment program. Value Health. 2010;13(8): 867-872.
6. Yang Y, Abel L, Buchanan J, Fanshawe T, Shinkins B. Use of decision modelling in economic evaluations of diagnostic tests: an appraisal and review of health technology assessments in the UK. PharmacoEconomics-open. 2019; 3(3): 281-291.
7. García-Mochón L, Balbino JE, de Labry Lima AO, et al. HTA and decision-making processes in Central, Eastern and South Eastern Europe: results from a survey. Health Policy. 2019; 123(2): 182-190.
8. Lampe K, Mäkelä M, Garrido MV, et al. The HTA core model: a novel method for producing and reporting health technology assessments. Int J Technol Assess Health Care. 2009; 25(S2):9-20.
9. Doctor J, MacEwan JP. Limitations of traditional health technology assessment methods and implications for the evaluation of novel therapies. Curr Med Res Opin. 2017; 33(9):1635-42.
10. Potter, BK, Avard D, Graham ID, et al. Guidance for considering ethical, legal, and social issues in health technology assessment: application to genetic screening. Int J Technol Assess Health Care. 2008; 24(4): 412.
11. Haverinen J, Keränen N, Falkenbach P, et al. Digi-HTA: Health technology assessment framework for digital healthcare services. FinJeHeW. 2019; 11(4): 326-341.
12. European Network for Health Technology Assessment (EUnetHTA). HTA Core Model Available from: https://eunethta.eu/hta-core-model. [Accessed Feb 26, 2020].
13. Jabareen Y. Building a conceptual framework: philosophy, definitions, and procedure. Int J Qual Methods. 2009;8(4):49-62.
14. European Commission. Next Generation Health Technology Assessment to support patient-centred, societally oriented, real-time decision-making on access and reimbursement for health technologies throughout Europe. Available from: https://cordis.europa.eu/project/id/825162. [Accessed March 20, 2020].
15. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467-73.
16. Chan K, Nam S, Evans B, et al. Developing a framework to incorporate real-world evidence in cancer drug funding decisions: the Canadian Real-world Evidence for Value of Cancer Drugs (CanREValue) collaboration. BMJ open. 2020;10(1):1-6.
17. Almeida N, Mines L, Nicolau I, et al. A Framework for Aiding the Translation of Scientific Evidence into Policy: The Experience of a Hospital-Based Technology Assessment Unit. Int J Technol Assess Health Care. 2019;35(3):204-11.

18. Angelis A, Kanavos P. Multiple criteria decision analysis (MCDA) for evaluating new medicines in health technology assessment and beyond: the Advance Value Framework. Soc Sci Med. 2017 ;188:137-56.
19. Veenstra DL, Roth JA, Garrison LP, Ramsey SD, Burke W. A formal risk-benefit framework for genomic tests: facilitating the appropriate translation of genomics into clinical practice. Genet Med. 2010;12(11):686-93.
20. Goetghebeur MM, Wagner M, Khoury H, et al. Combining multicriteria decision analysis, ethics and health technology assessment: applying the EVIDEM decisionmaking framework to growth hormone for Turner syndrome patients. Cost Eff Resour Alloc. 2010;8(1):4.
21. Neta G, Glasgow RE, Carpenter CR, et al. A framework for enhancing the value of research for dissemination and implementation. Am. J. Public Health. 2015; 105(1):49-57.
22. Graham ID, Logan J, Harrison MB, et al. Lost in knowledge translation: time for a map? J Contin Educ Health Prof. 2006;26(1):13-24.
23. Damschroder LJ, Aron DC, Keith RE, et al. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement. Sci. 2009;4(1):1-5.
24. Majdzadeh R, Sadighi J, Nejat S, Mahani AS, Gholami J. Knowledge translation for research utilization: design of a knowledge translation model at Tehran University of Medical Sciences. J Contin Educ Health Prof. 2008;28(4):270-7.
25. Newell WH, Wentworth J, Sebberson D. A theory of interdisciplinary studies. Issues in Interdisciplinary Studies. 2001.
26. Poulin P, Austen L, Scott CM, et al. Introduction of new technologies and decision making processes: a framework to adapt a local health technology decision support program for other local settings. Medical devices. 2013;6:185.
27. Tony M, Wagner M, Khoury H, et al. Bridging health technology assessment (HTA) with multicriteria decision analyses (MCDA): field testing of the EVIDEM framework for coverage decisions by a public payer in Canada. BMC Health Serv Res. 2011;11(1):329.
28. Brixner D, Kaló Z, Maniadakis N, Kim K, Wijaya K. An evidence framework for off-patent pharmaceutical review for health technology assessment in emerging markets. Value Health Reg Issues. 2018;16:9-13.
29. Miot J, Wagner M, Khoury H, Rindress D, Goetghebeur MM. Field testing of a multicriteria decision analysis (MCDA) framework for coverage of a screening test for cervical cancer in South Africa. Cost Eff Resour Alloc. 2012;10(1):2.
30. Abelson J, Wagner F, DeJean D, et al. Public and patient involvement in health technology assessment: a framework for action. Int J Technol Assess Health Care. 2016;32(4):256-64.
31. Assasi N, Tarride JE, O'Reilly D, Schwartz L. Steps toward improving ethical evaluation in health technology assessment: a proposed framework. BMC Med Ethics. 2016;17(1):34.
32. Gagnon MP, Desmartis M, Gagnon J, et al. Framework for user involvement in health technology assessment at the local level: Views of health managers, user representatives, and clinicians. Int J Technol Assess Health Care. 2015;31(1-2):68.
33. Hendricks S, Conrad N, Douglas TS, Mutsvangwa T. A modified stakeholder participation assessment framework for design thinking in health innovation. In Healthcare. 2018; 6(3): 191-196.
34. Rapport F, Clay-Williams R, Churruca K, et al. The struggle of translating science into action: foundational concepts of implementation science. J Eval Clin Pract. 2018;24(1):117-26.

35. Roberts JP, Fisher TR, Trowbridge MJ, Bent C. A design thinking framework for healthcare management and innovation. In Healthcare. 2016; 4(1): 1-14.

36. Globethics.net. Addressing Ethical and Moral Issues in Health Technology Assessment: Development of a Practical Framework. Report to the Canadian Centre for Ethics and corporate Policy Graduate Award Committee. Available from: https://52.208.232.29/handle/20.500.12424/19832. [Accessed Sep 26, 2021].

37. Ni M, Borsci S, Walne S, et al. The Lean and Agile Multi-dimensional Process (LAMP)–a new framework for rapid and iterative evidence generation to support health-care technology design and development. Expert Rev Med Devices. 2020;17(4):277-88.

38. Okumus F. Towards a strategy implementation framework. Int. J. Contemp. Hosp. Manag. 2001;13(7):327-338.

39. Meyers DC, Durlak JA, Wandersman A. The quality implementation framework: a synthesis of critical steps in the implementation process. Am J Community Psychol. 2012;50(3-4):462-80.

40. Vechakul J, Shrimali BP, Sandhu JS. Human-centered design as an approach for place-based innovation in public health: a case study from Oakland, California. Matern Child Health J. 2015;19(12):2552-9.

41. Brown T, Wyatt J. Design thinking for social innovation. Development Outreach. 2010.12(1):29-43.

42. Kilbourne AM, Neumann MS, Pincus HA, Bauer MS, Stall R. Implementing evidence-based interventions in health care: application of the replicating effective programs framework. Implementation Science. 2007;2(1):42.

43. Palozzi G, Brunelli S, Falivena C. Higher Sustainability and Lower Opportunistic Behaviour in Healthcare: A New Framework for Performing Hospital-Based Health Technology Assessment. Sustainability. 2018;10(10):3550.

44. Karnon J, Stahl J, Brennan A, Caro JJ, Mar J, Möller J. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–4. Med Decis Making. 2012;32(5):701-11.

45. Marsh K, IJzerman M, Thokala P, et al. Multiple criteria decision analysis for health care decision making—emerging good practices: report 2 of the ISPOR MCDA Emerging Good Practices Task Force. Value Health. 2016;19(2):125-137.

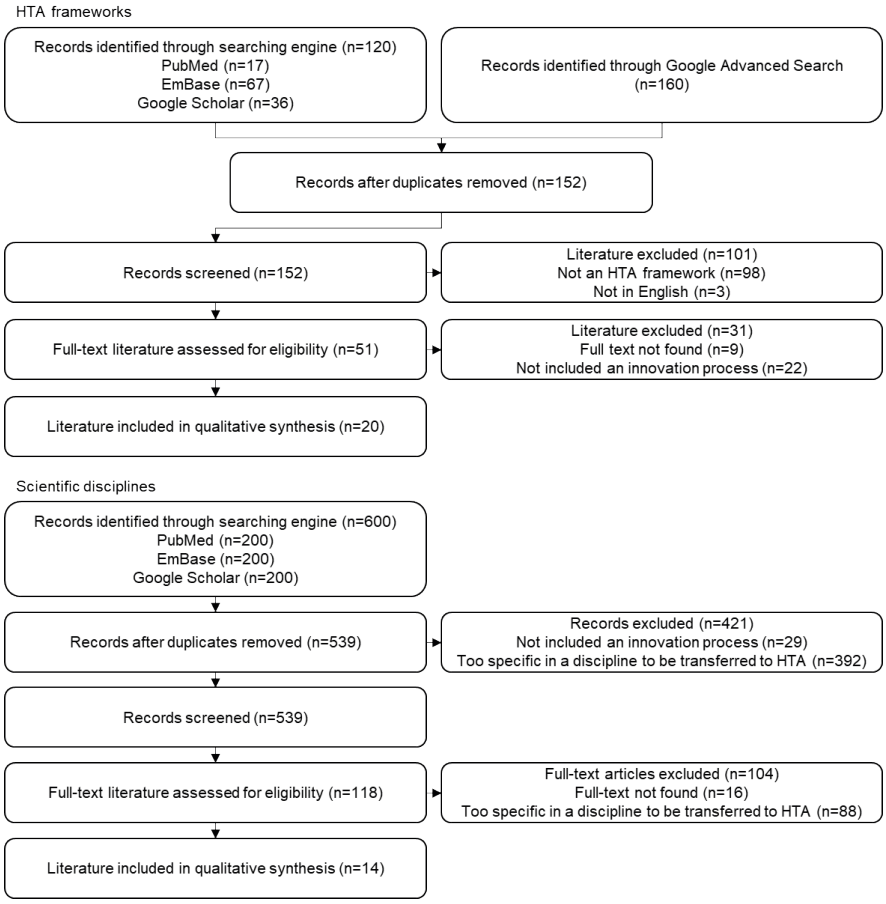
# Appendices

**Appendix 1. Search strategies of the scoping reviews for HTA frameworks and for scientific disciplines**

| Information sources | Codes | Filters |
|---|---|---|
| | ***Scientific Articles regarding HTA Frameworks*** | |
| PubMed | ("framework"[Title]) AND (("HTA"[Abstract] OR "health technology assessment"[Abstract]) OR ("HTA"[Title] OR "health technology assessment"[Title])) | • Article types: Journal Article<br>• Languages: English |
| Embase | 'framework':ti AND ('hta':ab,ti OR 'health technology assessment':ab,ti) | • Publication types: Article; Review |
| Google Scholar | allintitle: ("framework") AND ("HTA" OR "health technology assessment") | • Patents: not included<br>• Citations: included |
| | ***Grey Literature regarding HTA Frameworks*** | |
| • European Network for Health Technology Assessment (EUnetHTA)<br>• The Professional Society for Health Economics and Outcomes Research (ISPOR)<br>• Health Technology Assessment International (HTAi)<br>• The International Network of Agencies for Health Technology Assessment (INAHTA)<br>• Institute for Clinical and Economic Review (ICER)<br>• World Health Organization (WHO)<br>• National Institute for Health and Care Excellence (NICE)<br>• Google Advance Search | "HTA frameworks" | • Language: English<br>• Region: any region<br>• Terms appearing: anywhere in the page<br>• Safesearch: show most relevant results<br>• File type: any format<br>• Usage rights: not filtered by license |
| | ***Scientific Articles regarding Scientific Disciplines*** | |
| PubMed | ("framework"[Title] OR "model"[Title] OR "theory"[Title] OR "guidance") AND ("innovation"[Title] OR "identification"[Title] OR "research" [Title] OR "development"[Title] OR "implementation" [Title] OR "validation" [Title] OR "transferability" [Title] OR "generalization" [Title]) | • Text availability: Full Text<br>• Language: English<br>• Sort by Best Match |

**Appendix 1. Continued**

| Information sources | Codes | Filters |
|---|---|---|
| | ***Scientific Articles regarding Scientific Disciplines*** | |
| Embase | ("framework" OR "model" OR "theory" OR "guidance") AND ("innovation" OR "identification" OR "research" OR "development" OR "implementation" OR "validation" OR "transferability" OR "generalization") | • Search fields: Title<br>• Publication types: Article; Review<br>• Sort by Relevance |
| Google Scholar | allintitle: ("framework" OR "model" OR "theory" OR "guidance") AND ("innovation" OR "identification" OR "research" OR "development" OR "implementation" OR "validation" OR "transferability" OR "generalization") | • Citations: included<br>• Patents: not included<br>• Sort by Relevance |




**Appendix 2. Flow diagram of scanning and identifying eligible HTA frameworks and studies on scientific disciplines.**

**Appendix 3. Study characteristics of HTA frameworks and studies on scientific disciplines**

| HTA frameworks | | | | | |
|---|---|---|---|---|---|
| **Author** | **Year** | **Country (corresponding author)** | **Source** | **Journal name** | **Name of the HTA framework?** |
| Chan K et al. | 2020 | Canada | Scientific literature | BMJ Open | / |
| Ni M et al. | 2020 | The UK | Scientific literature | Expert Review of Medical Devices | The Lean and Agile Multi-dimensional Process (LAMP) |
| Almeida N et al. | 2019 | Canada | Scientific literature | International Journal of Technology Assessment in Health Care | / |
| Baran-kooiker A et al. | 2019 | Poland | Scientific literature | Acta Poloniae Pharmaceutica - Drug Research | The Evidence and Value: Impact on Decision Making Framework (EVIDEM) |
| Haverinen J et al. | 2019 | Finland | Scientific literature | Finnish Journal of eHealth and eWelfare | / |
| Brixner D et al. | 2018 | the USA | Scientific literature | Value in Health | / |
| Krahn M et al. | 2018 | Canada | Scientific literature | International Journal of Technology Assessment in Health Care | The Ontario Decision Framework |
| Palozzi G et al. | 2018 | Italy | Scientific literature | Sustainability | The Health Technology Balanced Assessment Framework (HTBA) |
| Angelis A et al. | 2017 | The UK | Scientific literature | Social Science & Medicine | The Advance Value Framework (AVF) |
| Assasi N et al. | 2016 | Canada | Scientific literature | BMC Medical Ethics | / |
| Abelson J et al. | 2016 | Canada | Scientific literature | International Journal of Technology Assessment in Health Care | / |

**Appendix 3. Continued**

| | | | HTA frameworks | | |
|---|---|---|---|---|---|
| **Author** | **Year** | **Country (corresponding author)** | **Source** | **Journal name** | **Name of the HTA framework?** |
| Gagnon M et al. | 2015 | Canada | Scientific literature | International Journal of Technology Assessment in Health Care | / |
| Widrig D et al. | 2014 | Switzerland | Scientific literature | International Journal of Technology Assessment in Health Care | / |
| Assasi N et al. | 2013 | Canada | Grey literature (Google advanced search) | / | / |
| Poulin P et al. | 2013 | Canada | Scientific literature | Medical Devices: Evidence and Research | / |
| Goetghebeur M et al. | 2012 | Canada | Scientific literature | Medical Decision Making | The Evidence and Value: Impact on Decision Making Framework (EVIDEM) |
| Miot J et al. | 2012 | South Africa | Scientific literature | Cost Effectiveness and Resource Allocation | The Evidence and Value: Impact on Decision Making Framework (EVIDEM) |
| Tony M et al. | 2011 | Canada | Scientific literature | BMC Health Services Research | The Evidence and Value: Impact on Decision Making Framework (EVIDEM) |
| Goetghebeur M et al. | 2010 | Canada | Scientific literature | Cost Effectiveness and Resource Allocation | The Evidence and Value: Impact on Decision Making Framework (EVIDEM) |
| Veenstra D et al. | 2010 | The USA | Scientific literature | Genetics in Medicine | / |

**Appendix 3. Continued**

| Scientific disciplines | | | | | |
|---|---|---|---|---|---|
| **Author** | **Year** | **Country (Corresponding author)** | **Source** | **Journal name** | **Scientific discipline categorization** |
| Hendricks S et al. | 2018 | South Africa | Scientific literature | Healthcare | Design thinking |
| Rapport F et al. | 2017 | Australia | Scientific literature | Journal of Evaluation in Clinical Practice | Implementation research |
| Neta G et al. | 2015 | The USA | Scientific literature | American Journal of Public Health | Implementation research |
| Roberts J et al. | 2015 | The USA | Scientific literature | Healthcare | Design thinking |
| Vechakul J et al. | 2015 | The USA | Scientific literature | Maternal Child Health Journal | Design thinking |
| Meyers D et al. | 2012 | The USA | Scientific literature | Am J Community Psychol | Implementation research |
| Brown T et al. | 2010 | The USA | Scientific literature | Stanford Social Innovation Review | Design thinking |
| Damschroder L et al. | 2009 | The USA | Scientific literature | Implementation Science | Implementation research |
| Liyanage C et al. | 2009 | The UK | Scientific literature | Journal of Knowledge Management | Implementation research |
| Majdzadeh R et al. | 2008 | Iran | Scientific literature | Journal of Continuing Education in the Health Professions | Implementation research |
| Kilbourne A et al. | 2007 | The USA | Scientific literature | Implementation Science | Implementation research |
| Graham I et al. | 2006 | Canada | Scientific literature | Journal of Continuing Education in the Health Professions | Implementation research |
| Newell W et al. | 2001 | The USA | Scientific literature | Issues in Integrative Studies | Interdisciplinary research |
| Okumus F et al. | 2001 | Turkey | Scientific literature | International Journal of Contemporary Hospitality Management | Implementation research |

Chapter 3

# Roadmap to Innovation of HTA Methods (IHTAM): Insights from Three Case Studies of Quantitative Methods

Li Jiu, Junfeng Wang, Jan-Willem Versteeg, Yingying Zhang, Lifang Liu, Francisco Javier Somolinos-Simón, Jose Tapia-Galisteo, Gema García-Sáez, Milou A Hogervorst, Xinyu Li, Aukje K Mantel-Teeuwisse, Wim G Goettsch

*Submitted*

# Abstract

## Background

A conceptual framework, called Innovation of Health Technology Assessment Methods (IHTAM), has been developed, to facilitate understanding of how to innovate methods of health technology assessment (HTA). However, the framework has not been validated in practice. Hence, we aimed to explore framework validity in three cases of method innovation that are part of the HTx project, and to develop a roadmap to improve framework applicability.

## Method

The IHTAM framework was applied to three cases of innovating HTA methods. We collected feedback from case study leaders and consortium members after a training session, an approximately one-year follow-up of periodic case study meetings, and a general assembly meeting where innovation progresses of the three cases were reported, through surveys and interviews. Feedback was then summarized using an open-coding technique.

## Results

According to feedback, the framework provided a structural way of deliberation and helped to improve collaboration among HTA stakeholders. However, framework applicability could be improved if it was complemented by a roadmap with a loop structure to provide tailored guidance for different cases, and with items to elaborate actions to be taken by stakeholders. Accordingly, a forty-eight-item roadmap was developed.

## Conclusions

The IHTAM framework was generally applicable to the three case studies. A roadmap, with loop structure and actionable items, could complement the framework, and may provide HTA stakeholders with tailored guidance on developing new methods. To further validate the framework, we recommend stakeholders to apply the IHTAM framework and its roadmap in future practice.

## Introduction

Methods of health technology assessment (HTA) refer to methods relevant to the full scope of an HTA process (1,2). According to the HTA Core Model from the European network for HTA (EUnetHTA), the HTA scope can be categorized into nine domains, including but not limited to clinical effectiveness, costs and economic evaluation, and patient and social aspects (3). Also, according to the European Patients' Academy on Therapeutic Innovation (EUPATI), an HTA process generally has three phases: collecting and reviewing scientific evidence of a health technology; making decisions on reimbursement and pricing; and implementing decisions and monitoring impact (4). Therefore, the term "HTA methods" has broad implications with a large number of examples. One example is the measurement of patient reported outcomes (PRO), through which patient aspects are considered during collection of evidence, such as quality of life (5). Another example is the use of decision-analytic models for health economic evaluation, which investigate clinical effectiveness and costs for HTA decision-making (6).

HTA methods may be repeatedly developed and implemented, in other words, innovated, for multiple reasons. One reason is the emergence of novel health technologies to which traditional HTA methods may not be suited. For example, complex health technologies, which include combinations of health technologies, personalized treatment, or treatment pathways, pose requirements for novel methods that support more tailored decision-making (7). Another reason is the changed availability of data that could be used for HTA. For example, the increasing use of real-world data (RWD) poses challenges on data quality, and creates needs for methods to assess quality of data sources (e.g. data registry) or studies using RWD (8). In addition, the variety of HTA settings (e.g. developed vs. developing countries) creates barriers to transferring an existing HTA method in one setting to another, and creates needs for improving existing methods or developing a new method in the local setting (9,10).

While innovation of HTA methods is often needed, HTA stakeholders, such as clinicians, policy-makers, patient associations, third-party payers, and healthcare industry (11), often lack a general understanding on how to innovate HTA methods, and how they could engage in the innovation process. To facilitate such understanding, a conceptual framework, called Innovation of HTA Methods (IHTAM), has been developed under the umbrella of large H2020 project, HTx, that is focused on the development of new HTA methods. The IHTAM framework is based on two scoping reviews and stakeholder inputs through surveys and iterative brainstorm sessions (1). The framework defines

a general innovation process with three phases (i.e. “Identification,” “Development,” and “Implementation”) and nine subphases (e.g. “Design Prototypes” and “Plan for Implementation”). Also, the framework illustrates how stakeholders could be involved, by clarifying three roles they can play (i.e. “Developers,” “Practitioners,” and “Beneficiaries”). An overview of the framework is shown in Appendix 1.

While the IHTAM framework was developed, it has not been validated in innovation practice of HTA methods. Also, the framework may have limitations that bring concerns to its applicability. For example, the three-phase innovation process of the IHTAM framework was partly coded from innovation processes of existing HTA methods from a scoping review, and these methods were frameworks, such as those to incorporate real-world evidence, value-based criteria, or patient inputs in decision-making (12-14), rather than other types of methods (e.g. cost-effectiveness models).

Hence, the aim of this study was to explore applicability of the IHTAM framework in three cases of development of quantitative methods, and to improve applicability by updating the framework. This research was performed as part of the HTx project. The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825162 (15).

# Methods

## Case description

Three cases were identified from the HTx project based on the four case studies that are the fundament of HTx (15). The first case (CS1) involved innovation of models to evaluate cost-effectiveness of normal tissue complication probability (NTCP) models in head and neck cancer patients who are treated with protontherapy. NTCP models are models used in the field of radiotherapy to estimate the risk (i.e. probability of occurring) (16) of radiation-induced complications (17). The second case (CS2) involved innovation of models to predict risk of complications in patients with type-1 or type-2 diabetes. The complications refer to macrovascular complications (e.g. coronary heart disease), microvascular complications (e.g. diabetic renal disease), and short-term complications (e.g. hyperglycemia). The third case (CS3), related to the innovation of methods to use RWD in HTA settings. The methods included target trial emulation (TTE), longitudinal targeted maximum likelihood estimation (LTMLE), and causal machine learning (CML). TTE is a method to apply the study design principles of randomized trials to observational studies that aim to estimate the causal effect of an intervention (18,19). TMLE is a method to estimate causal effects using observational data (20). CML are

machine learning models that involve the process of identify causal inference (21). While these methods have been developed, they were not readily applicable to HTA, due to quality concerns (e.g. time-varying confounding) (22,23).

## Application of the IHTAM framework

According to the IHTAM framework (1), case study leaders were recommended to apply the framework in the following steps: First, to consider all (sub)phases and three innovation roles (i.e. developer, practitioner, and beneficiary) within the IHTAM framework, and to judge their relevance to the case; Second, to discuss whether new (sub)phases or roles of innovation apply; Third, to consider challenges of innovation; and Forth, to facilitate collaboration by inviting HTA stakeholders from multiple backgrounds (e.g. patients and industry), based on the case-specific needs. To ensure case study leaders understood how to apply the framework, we followed up the framework application progress in each case, and provided assistance in three steps.



In the beginning, we organized a face-to-face training session for case study leaders and consortium members during the HTx project consortium meeting, in April, 2022. During the training session, one researcher (LJ) introduced the structure of the IHTAM framework, and explained how to apply the framework, using the patient-reported-outcome-measures (PROM) toolbox, co-developed earlier as part of the HTx project, as an example method (24). Any confusion from stakeholders was solved through questions and answers.

Next, we followed up each case, by attending the regular meetings, which were held approximately every two months for each case. In each meeting, at least one researcher (JW or LJ) attended, reminded case study leaders to keep applying the IHTAM framework, and answered relevant questions. The follow-up lasted for one year.

At the end of the follow-up, case study leaders were asked to systematically report the method innovation progress, using the IHTAM framework. Each case was given 30 minutes during the face-to-face general assembly meeting of the HTx project, in May 2023. Before the general assembly meeting, we provided case study leaders with a slide template, and resolved any outstanding questions through e-mail or an online meeting.

## Evaluation & Improvement of the framework applicability

After the general assembly meeting in 2023, we invited case study leaders and consortium members to provide feedback on the applicability of the IHTAM framework, through an online survey or interview, based on the invitees' preference.

The online survey or interview involved two open questions: First, which aspect of the IHTAM framework could improve the understanding of progress made in case studies, and second, which aspect of the IHTAM framework did not improve the understanding and could be further improved. All feedback was recorded by one researcher (LJ), and then independently summarized by two researchers (LJ and JV), using NVIVO12. Any discrepancy was solved through discussion.

Based on the feedback, we updated the IHTAM framework to improve its applicability. The updated version was first prepared by authors, then edited by case study leaders and consortium members, and finalized by four authors (LJ, JV, AM, and WG) in a group meeting.

# Results

## Framework application to case studies

After the one-year follow-up (about six periodic meetings for each case), three case study leaders, who were researchers, and twenty-four consortium members, from research institutes (n=20), HTA agencies (n=4), and patient organizations (n=1) attended the general assembly meeting. The general progress of HTA method innovation, reported by case study leaders, is shown in Figure 1.

In summary, the "Implementation" phase included one or two subphases (e.g. "Learn from Past & Present") that were in planning by at least one of the case studies. Additionally, none of the case studies completed all (sub)phases, but CS3 made the plans for the subphases yet to be conducted. In the "Identification" phase, some gaps in the HTA field and limitations of existing methods were identified, from literature reviews (all cases) or by observing practice in HTA settings (CS3). For example, in CS1, the proton therapy had clinical benefits, but its high economic burden and low capacity restricted its access to patients with head and neck cancer. While the NTCP models used to select patients for proton therapies were available, information was lacking on the cost-effectiveness of these models. In the subphase "Imagine Future", scenarios on what future HTA processes may look like were identified through feedback obtained during periodic HTx meetings (CS2) or workshops of HTA agencies plus a scoping review (CS3). In contrast, in CS1, a plan was made to identify future scenarios for using NTCP models for treatment and reimbursement decision-making.

According to the previous two subphases, the needs of the novel HTA method(s) were identified in the three cases. In the "Development" phase, key resources needed for developing a method were gathered in all the three cases (subphase "Manage Resources for Innovation"). More specifically, all cases involved the collection of data used for method development (e.g. cancer registry data in CS1), while CS1 and CS3 reported case-specific resources. For example, in CS3, experts in the field of machine learning and decision-modelling, a case-specific human resource, were invited to aid with method development. After resource management, all cases involved a case-specific process of developing method prototypes. For example, CS2 involved cluster analyses and development of risk prediction models using machine learning techniques, and CS3 needed clinicians' inputs for method development. In the subphase "Pilot testing", sensitivity or scenario analyses were conducted in CS1 and CS2, to investigate method uncertainty or performance. In contrast, a plan was made in CS3 on this subphase, and data matching some HTA contexts would be used to test the methods in the future.

In the "Implementation" phase, all the cases involved a case-specific process of planning for implementation. In CS1, a workshop was organized to disseminate the method (i.e. a cost-effectiveness model), and to explain how the method was linked to the HTA decision-making policy in Europe. In CS3, several workshops were organized to not only disseminate the methods but also understand the motivations of potential case-specific practitioners (e.g. HTA agencies) or beneficiaries (e.g. clinicians) to adopt the methods. In contrast, method dissemination in CS2 was conducted through developing a tool (i.e. decision-support tool) for potential model users (e.g. clinicians). Additionally, a plan of model external validation was made in CS2 to investigate model transferability across countries. In the subphase "Apply a Method to Practice", only CS2 reported ongoing tasks, as risk prediction models were being incorporated into a cost-effectiveness model, and relevant patient subgroups were being applied to HTA cluster analyses. While CS3 involved no ongoing task in this subphase, a plan was made by developers to provide technical assistance to future practitioners who feel interested. Lastly, no case involved ongoing tasks related to the subphase "Test & Transfer", which involved testing method performance during implementation with the intention of further innovation. While all case studies had not entered this subphase, leaders of CS3 considered it relevant and planned for externally validating methods they developed.

## Evaluation of the IHTAM framework applicability

Of the twenty-eight attendees of the general assembly meeting in May 2023, when the case progresses were reported using the IHTAM framework, three were authors who collected feedback and updated the framework, seven were case study leaders, and eighteen were consortium members. Two case study leaders (labelled as L1-L2)

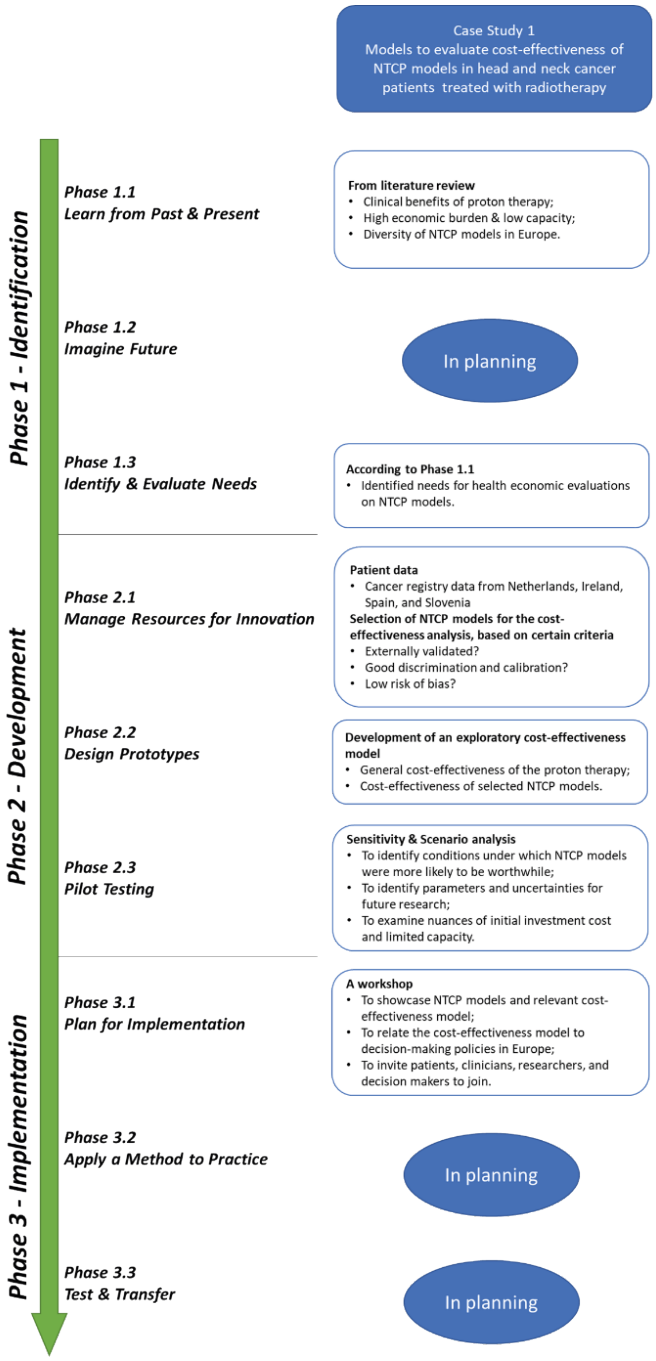
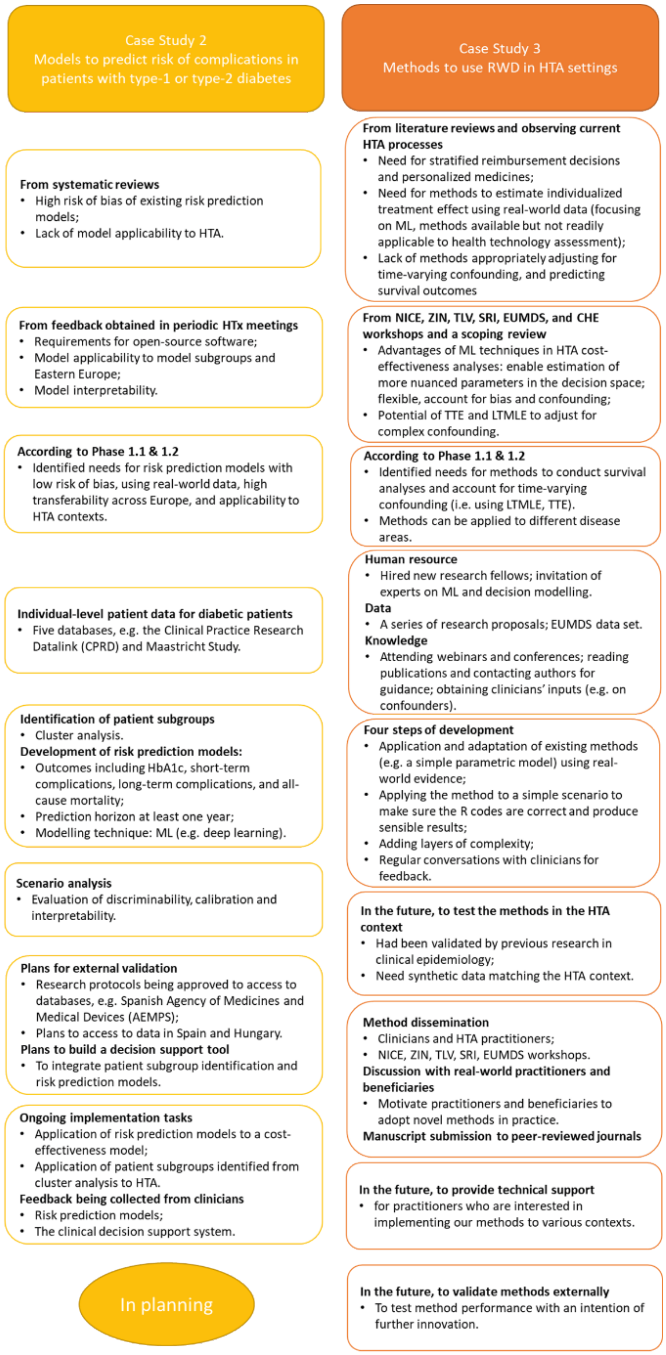



**Figure 1. General innovation progress of HTA methods, reported by leaders of the three case studies.**

ML indicates machine learning; NTCP, normal tissue complication probability; RWD, real-world data; HTA, health technology assessment; TTE, target trial emulation; LTMLE, longitudinal targeted maximum likelihood estimation; NICE, National Institute for Health and Care Excellence (England); ZIN, National




Health Care Institute (Netherlands); TLV, Dental and Pharmaceutical Benefits Agency (Sweden); SRI, Syreon Research Institute (Hungary); CHE, Center for Health Economics (University of York); EUMDS, European Myelodysplastic Syndromes Registry.

and 12 consortium members (labelled as A1-A12) provided feedback. Of those without feedback, only one consortium member mentioned the reason: "did not realize that case study leaders tried to use the IHTAM framework", so the framework could not be judged. In summary, all case study leaders and most consortium members (n=8) who provided feedback stated that, the IHTAM framework had improved their understanding of HTA methods innovation. Meanwhile, all case study leaders and most consortium members (n=7) pointed out current limitations of the framework and provided suggestions on how to address them.

The improved understanding of consortium members could mainly be summarized into three points. First, two case study leaders and six consortium members mentioned that the IHTAM framework provided a structural way of thinking, and it helped avoid neglecting some important innovation (sub)phases. For example, "the framework helps to structure discussions about the case studies", "standardizes' our ontology", and shows "how the next element in a study builds upon the previously completed tasks" (A2 & A8). Also, "the IHTAM framework is "well-designed and provides guidance of good practice" (A11). Second, all case study leaders and five consortium members mentioned that the framework was relevant to the innovation process of case studies. For example, "stakeholders could obtain many details on the needs for an HTA method and how it was developed" (L1). Also, the Implementation phase is "practical", as it evaluates "where is a capacity to apply methods" and "where the healthcare system can benefit from it" (A3). Lastly, all case study leaders and five consortium members mentioned that the framework could improve multi-multidisciplinary collaboration, as it "could be understood by stakeholders without any HTA background" (A4) and reminds stakeholders that "innovative methods need more collaborative efforts" (A12).

The framework limitations could be mainly summarized into four points. First, as mentioned by four consortium members, the framework might need a loop structure. For example, it could include a "spiral" structure to include "long learning circles" of innovation (A2), or it was not necessary to "move to clockwise direction" (A1). Second, a checkbox (template, checklist, etc.) attached to the conceptual framework might increase user-friendliness. The reason was that stakeholders could "know what has been done and what could be done in the future" (L1), and that, it was "helpful for users to report it to audience" (A6). Third, stakeholders "might have different understanding of what each step means to them" (A11), and could "misinterpret" some of these steps, e.g., whether "'Design Prototypes' included modelling" (A12). To avoid misinterpretation, one case study leader and four consortium members expressed the need for a tool to complement the conceptual framework, which could provide further guidance and elaboration, and preferably, "incorporate tips for different stakeholders" (A6).

Lastly, four case study leaders and three consortium members stated that, some IHTAM subphases were not yet conducted, and only a plan was made. For example, " 'Pilot testing' is hard to follow in the case of developing risk prediction models" (L2). Similarly, a case study leader (A11) thought "some steps do not apply" or "skippable", as they depended on whether a method was developed or only transferred to another HTA context.

### A roadmap to complement the conceptual framework



A roadmap was designed to complement the IHTAM framework, taking the identified limitations into account. A flow diagram and snapshot of the roadmap is shown in Figure 2, and the details are shown in Appendix 2. The roadmap has three main features. First, it includes forty-eight items that covers all content of the original conceptual framework. With the roadmap, HTA stakeholders can more easily know what has been done and what to do next, by simply comparing roadmap items with their actions. Second, the roadmap includes two types of items, which interrelate each other. Action items show what actions of innovation may need to be done, while Reporting items show issues to be reported to the audience (i.e. Reporting items). With the two items, some issues around misinterpretation can be solved, as sufficient reporting of actions can help avoid misunderstandings of stakeholders from various knowledge backgrounds. The third feature of the roadmap is a loop structure, which enables the design of a case-specific innovation process. Under each Action item, stakeholders are asked to judge whether the action has been taken in their case. Based on the judgement, the roadmap leads stakeholders to different items. With the loops, (sub)phases or actions considered irrelevant to a case can be skipped, while those considered relevant can even be repeatedly conducted.

## Discussion

In this study, we applied the Innovation of HTA Methods (IHTAM) framework to three case studies in the HTx project, which were relevant to innovating a cost-effectiveness model, risk prediction models, and approaches to exploiting real-world evidence in HTA settings. The IHTAM framework was in general appreciated in the three case studies, as it provided a structural way of thinking, is highly relevant to the innovation process of case studies, and it could improve multi-multidisciplinary collaboration. Based on feedback from case study leaders and consortium members of the HTx project who were informed on the reports of the three cases, we developed a roadmap, which could complement the original conceptual framework by overcoming its limitations.

The IHTAM framework complemented by the roadmap could add value to HTA stakeholders who are involved in HTA method innovation. First, it facilitates knowledge

transfer and exchange (KTE) among stakeholders with different knowledge backgrounds. KTE is an interactive process, and one of its primary purposes is to increase the likelihood that research evidence will be used in policy and practice decisions (25). In HTA settings, where HTA methods are applied to clinical or reimbursement decision-making, KTE is considered difficult as it involves a series of complex actions (26). Our roadmap may partly solve the complexity, as it can awaken the realization of a knowledge gap, between HTA stakeholders who are already involved in innovation and those who are not yet involved. For example, according to our roadmap, some items that need to be reported during method development include: how the versions of a method prototype could address the heterogeneous needs (Item R2.2.1), and how the ease of implementation was considered in the method prototype need to be specified (Item R2.2.2). While developers are familiar with how methods they develop can be used, reporting such information to beneficiaries, who are not directly involved in method innovation but could benefit from a novel method, could enhance beneficiaries' motivation. More specifically, beneficiaries, who are policy makers in some cases, may explore added value of a novel method to HTA regulations in an early stage, and to provide in-time feedback that improves method transferability.

Another advantage of the roadmap is that it may motivate HTA stakeholders to participate in the action of method innovation. Given the quite long circle of a whole HTA method innovation process, a single stakeholder, regardless of the innovation role (e.g. developer), hardly participates in the whole process. As shown in our study, in the end of the one-year follow-up, none of the three case studies went through all innovation (sub)phases, and only CS2 started to apply methods to practice. This long-circle challenge could be addressed by the updated framework. More specifically, by quantifying (sub)phases of the IHTAM framework into actionable items, stakeholders may know where they should take responsibility and when their roles may be taken over. For example, in CS2, some practitioners such as HTA modelers, applied the risk prediction models to a cost-effectiveness analysis. According to the IHTAM framework and roadmap, practitioners could describe how model developers are approached, and describe the types of assistance they need (e.g. how to load the risk prediction models in another software) (Item A3.2.1 & R3.2.1). Developers of risk prediction models could record feedback from practitioners after a cost-effectiveness model is developed (Item A3.2.2 & R3.2.2).

We recommend HTA stakeholders to use the roadmap in four steps. First, stakeholders may scan the IHTAM framework and the roadmap, to understand how to innovate an HTA method and how to involve other stakeholders in innovation. Second, stakeholders may understand the current innovation status of their cases and identify corresponding items in the roadmap as their starting point. The innovation does not necessarily start

from identifying limitations of existing methods (i.e. Item A1.1.1), but it could start from any IHTAM (sub)phase of item. Moreover, an innovation process is not necessarily initiated by developers. For example, during the implementation of an existing method, practitioners may sense a lack of method transferability to a certain context. Then they could start the innovation loop from the IHTAM subphase "Test & Transfer", or the Item A3.3.1 of the roadmap. After evaluating the validity of methods during innovation and method adoption of practitioners, they could make a decision on whether to initiate another round of identifying limitations of existing HTA methods (Item A3.3.3 & A3.3.4).

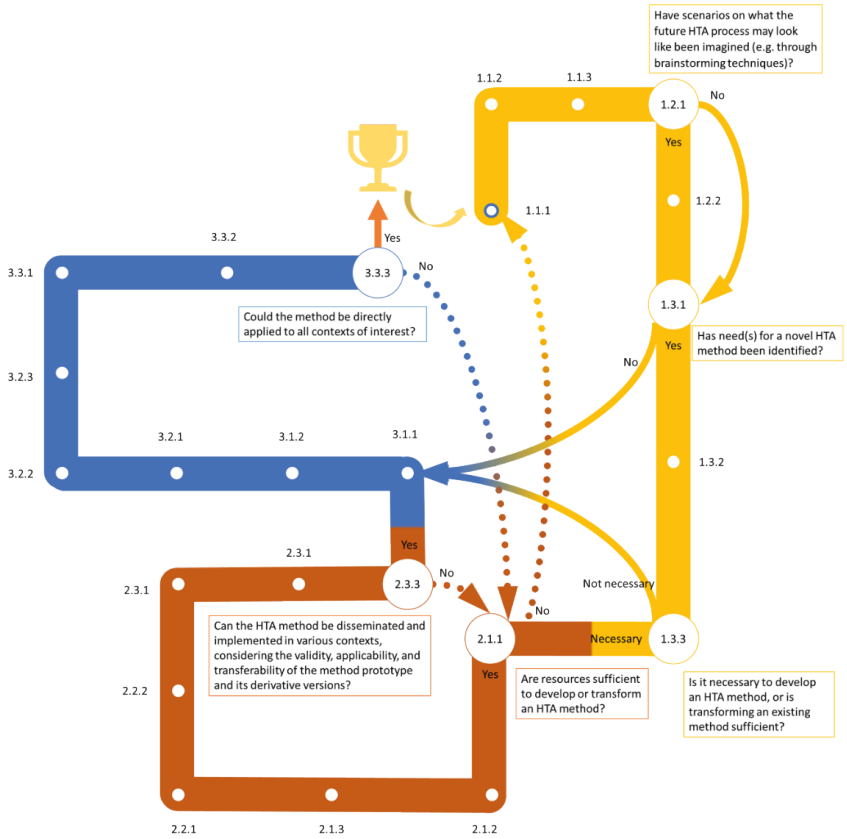



**Figure 2. A flow diagram to illustrate the IHTAM roadmap.**
The three-digit numbers indicate the roadmap items: the first digit indicates the three innovation phases (i.e. "Identification", "Development", "Implementation"), which are colored in yellow, red, and blue, respectively; the second digit, the nine subphases (e.g. "Learn from past & present" and "Imagine future"); the third digit, the specific items of a (sub)phase. The hollow circles indicate the items linked to the loop structure: e.g., if needs for a novel HTA method is identified (Item 1.3.1), users may manage resources needed for method development (Item 1.3.2); otherwise, user may jump to Item 3.1.1, to plan for implementation of an existing method. The arrows indicate the loop structure: solid arrows, going forward; dashed arrows, going back.

Once a starting point is identified, stakeholders may clarify their roles (e.g. developers, practitioners, and beneficiaries) and divide tasks accordingly. Moreover, an individual stakeholder may determine its stopping point, and if feasible, propose to stakeholders who to take over. One practical way of determining the stopping point is to draw a timeline for relevant tasks, based on task magnitude, and to assign the involved stakeholders.

Lastly, it is recommended to build a log of innovation throughout an innovation process by following the IHTAM framework and its roadmap. The innovation log could record the actions of innovation and all relevant details. The innovation log could help stakeholders who participate afterwards view the landscape and understand details that are relevant to their roles. Current research projects, with a goal to innovate HTA methods, have already recorded innovation progresses in some ways. Still, with an innovation log, HTA stakeholders could take a step further, to link all relevant documents, and to help themselves figure out their roles in a big research project.

Our study has a number of limitations. One limitation is that, only consortium members within the HTx project provided feedback regarding framework applicability, and only half of those responded. Another limitation is that more than half of the consortium members were more or less involved in at least one case study, so they had prior information (though this maybe not complete) on case studies before they responded to reports from case study leaders. The above-mentioned limitations could cause an overestimation of model applicability. Still, as two researchers independently summarized feedback with the rigid coding technique, the obtained feedback could objectively reflect the limitations of the conceptual framework approach. As the roadmap complementing the IHTAM framework was developed, we believe that the applicability of the updated framework is considerably improved. Also, to further test the framework applicability, we recommend HTA stakeholders with various backgrounds (e.g. payers and industry) outside the HTx project to apply the framework to the innovation processes of different types of methods. Another limitation is that we only applied the IHTAM framework to three cases of quantitative methods, e.g., models, rather than qualitative methods. However, we believe the stakeholder's feedback on framework applicability, as well as the designed roadmap, is transferable to qualitative methods. One reason is that the IHTAM framework was originally developed based on actual innovation processes of qualitative methods, and the roadmap enables a method-specific innovation process.

## Conclusions

The IHTAM framework was generally applicable to case studies of innovating HTA methods. A roadmap and the conceptual framework approach could help facilitate knowledge transfer and exchange among HTA stakeholders with different knowledge backgrounds. Also, it could motivate stakeholders from understanding method innovation to action. To further validate the framework, we recommend HTA stakeholders with various backgrounds (e.g. payers and industry) outside the HTx project to apply the framework to the innovation processes of different types of methods.



## Author contribution

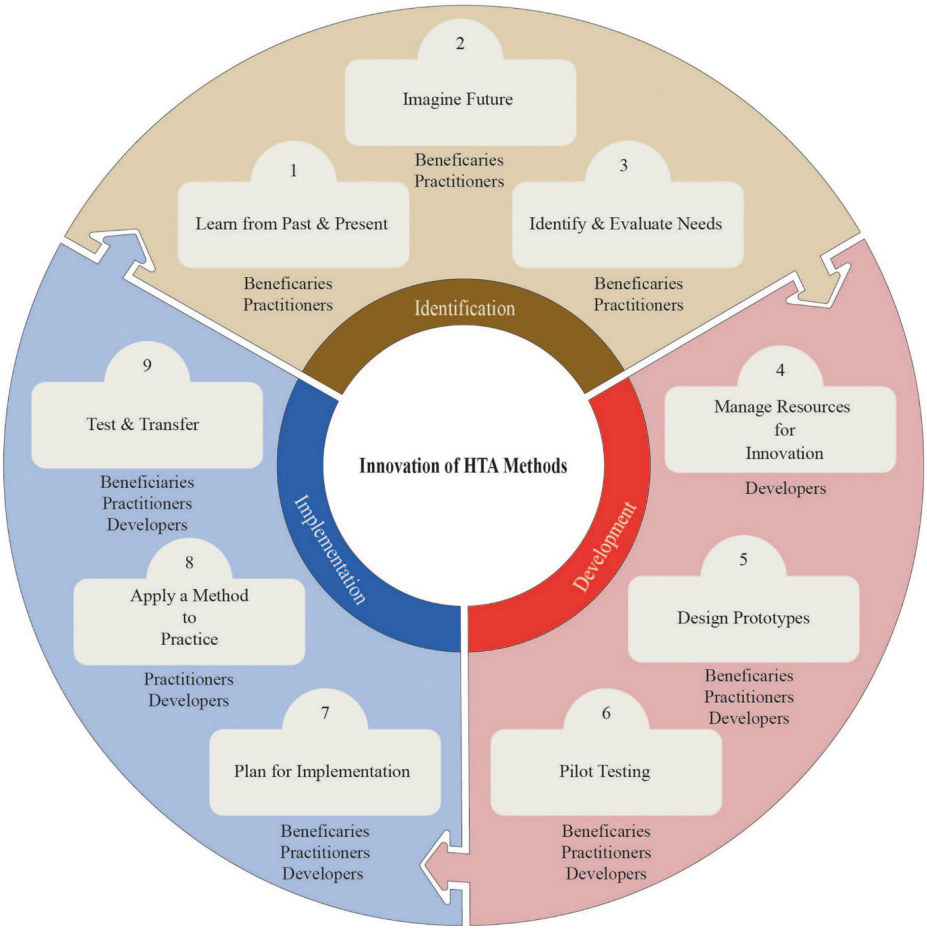
LJ organized the training session for applying the conceptual framework in the case studies, followed up the case studies by attending the regular meetings, obtained and synthesized feedback from stakeholders within the HTx project, edited the conceptual framework, and wrote the draft manuscript. All co-authors contributed to this study and critically reviewed and approved the manuscript.

# References


1. Jiu L, Hogervorst MA, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. Understanding innovation of health technology assessment methods: the IHTAM framework. Int J Technol Assess Health Care. 2022;38(1).
2. World Health Organization (WHO). Health technology assessment glossaries. Available from: https://www.who.int/health-technology-assessment/about/Glossaries/en. [Accessed Jun 28, 2023].
3. Kristensen FB, Lampe K, Wild C, Cerbo M, Goettsch W, Becla L. The HTA Core Model®—10 years of developing an international framework to share multidimensional value assessment. Value Health. 2017 Feb 1;20(2):244-50.
4. The European Patients' Academy (EUPATI). Health Technology Assessment process: Fundamentals. Available from: https://toolbox.eupati.eu/resources/health-technology-assessment-process-fundamentals. [Accessed Jun 28, 2023].
5. Brettschneider C, Lühmann D, Raspe H. Informative value of patient reported outcomes (PRO) in health technology assessment (HTA). GMS Health Technol Assess. 2011;7.
6. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess. 2004 Jan 1;8(36):iii-v.
7. Hogervorst MA, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. Reported challenges in health technology assessment of complex health technologies. Value Health. 2022 Jun 1;25(6):992-1001.
8. Oortwijn W, Sampietro-Colom L, Trowman R. How to deal with the inevitable: generating real-world data and using real-world evidence for HTA purposes–from theory to action. Int J Technol Assess Health Care. 2019;35(4):346-50.
9. Lou J, Sarin KC, Toh KY, Dabak S, Adler A, Ahn J, et al. Real-world data for health technology assessment for reimbursement decisions in Asia: current landscape and a way forward. Int J Technol Assess Health Care. 2020 Oct;36(5):474-80.
10. Garrido MV, Gerhardus A, Røttingen JA, Busse R. Developing health technology assessment to address health care system needs. Health Policy (New York). 2010 Mar 1;94(3):196-202.
11. Cavazza M, Jommi C. Stakeholders involvement by HTA Organisations: Why is so different?. Health Policy (New York). 2012 May 1;105(2-3):236-45.
12. Chan K, Nam S, Evans B, de Oliveira C, Chambers A, Gavura S, et al. Developing a framework to incorporate real-world evidence in cancer drug funding decisions: the Canadian Real-world Evidence for Value of Cancer Drugs (CanREValue) collaboration. BMJ Open. 2020 Jan 1;10(1):e032884.
13. Krahn M, Miller F, Bayoumi A, Brooker AS, Wagner F, Winsor S, et al. Development of the Ontario decision framework: a values based framework for health technology assessment. Int J Technol Assess Health Care. 2018;34(3):290-9.
14. Abelson J, Wagner F, DeJean D, Boesveld S, Gauvin FP, Bean S, et al. Public and patient involvement in health technology assessment: a framework for action. Int J Technol Assess Health Care. 2016;32(4):256-64.
15. HTx. About HTx project. Available from: https://www.htx-h2020.eu/about-htx-project. [Accessed June 28, 2023].
16. Grant SW, Collins GS, Nashef SA. Statistical Primer: developing and validating a risk prediction model. Eur J Cardiothorac Surg. 2018 Aug 1;54(2):203-8.

17. Van den Bosch L, Schuit E, van der Laan HP, Reitsma JB, Moons KG, Steenbakkers RJ, et al. Key challenges in normal tissue complication probability model development and validation: towards a comprehensive strategy. Radiother Oncol. 2020 Jul 1;148:151-6.
18. Frey BB, editor. The SAGE encyclopedia of educational research, measurement, and evaluation. New York: Sage Publications; 2018 Jan 29.
19. Matthews AA, Danaei G, Islam N, Kurth T. Target trial emulation: applying principles of randomised trials to observational studies. BMJ. 2022 Aug 30;378.
20. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. Am J Epidemiol. 2017 Jan 1;185(1):65-73.
21. Sanchez P, Voisey JP, Xia T, Watson HI, O'Neil AQ, Tsaftaris SA. Causal machine learning for healthcare and precision medicine. R Soc Open Sci. 2022 Aug 3;9(8):220638.
22. Gomes M, Latimer N, Soares M, Dias S, Baio G, Freemantle N, et al. Target trial emulation for transparent and robust estimation of treatment effects for health technology assessment using real-world data: opportunities and challenges. Pharmacoeconomics. 2022 Jun;40(6):577-86.
23. Crown WH. Real-world evidence, causal inference, and machine learning. Value Health. 2019 May 1;22(5):587-92.
24. National Health Care Institute (ZIN). PROM toolbox (summary in English). Available from: https://www.zorginzicht.nl/ondersteuning/prom-toolbox-summary-in-english. [Accessed June 28, 2023]
25. Mitton C, Adair CE, McKenzie E, Patten SB, Perry BW. Knowledge transfer and exchange: review and synthesis of the literature. Milbank Q. 2007 Dec;85(4):729-68.
26. Beretta V. Development and Implementation of Health Technology Assessment: Turning Knowledge Into Action. Berlin: Springer Nature; 2021 Mar 27.

# Appendices




**Appendix 1. Concept map of the IHTAM framework.**

**Appendix 2. Roadmap of the IHTAM framework**

| Actions for innovating an HTA method | Reporting a process of innovation |
|---|---|
| ***Identification Phase - 1.1 - Learn from Past & Present*** | |
| A1.1.1 Has the research technique(s) (e.g. surveys, interviews, literature review) been used to identify limitations of an existing HTA method?<br>✓ Go to R1.1.1 & A1.1.2. | R1.1.1<br>✓ If Yes<br>Report the research technique(s) used.<br>✓ If No<br>Specify the reason why the research technique(s) is not used. |
| A1.1.2 Have all relevant HTA stakeholder groups been involved in identifying limitations of existing HTA methods?<br>✓ Go to R1.1.2 & A1.1.3. | R1.1.2<br>✓ If Yes or No<br>Report HTA stakeholders involved (e.g. HTA agencies, clinicians, patients, etc.).<br>✓ If No<br>Specify how the lack of stakeholder groups could influence identification of needs of a novel HTA method. |
| A1.1.3 Have limitations of existing HTA methods been identified?<br>✓ If Yes, go to R1.1.3 & A1.2.1.<br>✓ If No, go to A1.2.1. | R1.1.3<br>✓ If Yes<br>List the identified limitations of existing HTA methods; If available, specify the heterogeneity of limitations identified by different HTA stakeholders. |
| ***Identification Phase - 1.2 - Imagine Future*** | |
| A1.2.1 Have scenarios on what the future HTA process may look like been imagined (e.g. through brainstorming techniques)?<br>✓ If Yes, go to R1.2.1 & A1.2.2.<br>✓ If No, go to A1.3.1. | R1.2.1<br>✓ If Yes<br>Specify the technique(s) used to imagine future scenarios; Describe what future scenarios may look like.<br>✓ If No<br>Specify the reason why future scenarios is not imagined. |
| A1.2.2 Have enablers and barriers to reach future scenarios been outlined?<br>✓ If Yes, go to R1.2.2 & A1.3.1.<br>✓ If No, go to A1.3.1. | R1.2.2<br>✓ If Yes<br>List the technique(s) used to outline enablers and barriers, such as interviews and surveys; List enablers and barriers to reach future scenarios; Propose potential solutions to utilize facilitators and to overcome barriers. |
| ***Identification Phase - 1.3 - Identify & Evaluate Needs*** | |
| A1.3.1 Based on Phase 1.1 & 1.2, has need(s) for a novel HTA method been identified?<br>✓ If Yes, go to R1.3.1 & A1.3.2.<br>✓ If No, go to A3.1.1, to plan to implement an existing method. | R1.3.1<br>✓ If Yes<br>Define the need(s) to be satisfied by an HTA method; According to R1.1.3 & R1.2.1 & R1.2.2, summarize how the needs are identified. |
| ***Identification Phase - 1.3 - Identify & Evaluate Needs*** | |

**Appendix 2. Continued**

| Actions for innovating an HTA method | Reporting a process of innovation |
|---|---|
| A1.3.2 Does the need(s) vary across contexts (e.g. different types of health technologies, disease areas, geographic areas, etc.)?<br>✓ Go to R1.3.2 & A1.3.3. | R1.3.2<br>✓ If Yes or No<br>Define the context(s) where a novel HTA method is needed.<br>✓ If No<br>Specify why the need(s) can be similar across contexts. |
| A1.3.3 Is it necessary to develop an HTA method, or is transforming an existing method sufficient?<br>✓ If developing an HTA method is necessary, go to R1.3.3 & A2.1.1.<br>✓ If not necessary, go to R1.3.3 & A3.1.1, to evaluate transferability of an existing method. | R1.3.3<br>✓ If Yes or No<br>Specify the rationale to develop an HTA method or to transform an existing method. |
| ***Development Phase - 2.1 - Manage Resources for Innovation*** | |
| A2.1.1 Are resources sufficient to develop or transform an HTA method?<br>✓ If Yes, go to R2.1.1 & A2.1.2.<br>✓ If No, go to R2.1.1 & A1.1.1. | R2.1.1<br>✓ If Yes<br>Report the technique(s) used to evaluate the resource sufficiency; List all resources needed to develop or transform an HTA method, such as time, finance, and knowledge.<br>✓ If No<br>Specify why resources are not sufficient. |
| A2.1.2 Is it necessary to set priorities for needs (e.g. those identified from various contexts) that a method addresses, given limited resources?<br>✓ Go to R2.1.2 & A2.1.3. | R2.1.2<br>✓ If Yes or No<br>Specify the reason why setting priorities is (not) necessary.<br>✓ If Yes<br>Report the priority list for needs; Specify the rationale for the priority. |
| A2.1.3 Have HTA stakeholders other than researchers been invited for method development or transformation?<br>✓ Go to R2.1.3 & A2.2.1. | R2.1.3<br>✓ If Yes or No<br>List all HTA stakeholders involved in developing or transforming an HTA method.<br>✓If No<br>Specify the reasons why stakeholders other than researchers are not invited. |
| ***Development Phase - 2.2 - Design Prototypes*** | |
| A2.2.1 Have a method prototype and its derivative versions been developed or transformed, based on the heterogeneity of needs across contexts?<br>✓ Go to R2.2.1 & A2.2.2. | R2.2.1<br>✓ If Yes<br>Specify how the versions of a method prototype could address the heterogeneous needs.<br>✓ If No<br>Specify the reason why the heterogeneity of needs are not considered. |
| ***Development Phase - 2.2 - Design Prototypes*** | |

**Appendix 2. Continued**

| Actions for innovating an HTA method | Reporting a process of innovation |
|---|---|
| A2.2.2 Has the ease of implementing the method in practice been considered, when developing or transforming a method prototype?<br>✓ Go to R2.2.2 & A2.3.1. | R2.2.2<br>✓ If Yes<br>Specify how the ease of implementation was considered in the method prototype.<br>✓ If No<br>Specify how it could impact the method implementation. |
| ***Development Phase - 2.3 - Pilot Testing*** | |
| A2.3.1 Have pilot case studies been conducted to test validity of the method prototype and its derivative versions?<br>✓ Go to R2.3.1 & A2.3.2. | R2.3.1<br>✓ If Yes<br>Describe case studies conducted and HTA stakeholders involved; Evaluate how the case studies could simulate real-world practice; Report validity of the method prototype and its derivative versions.<br>✓ If No<br>Specify how it could impact the method validity. |
| A2.3.2 Has applicability to other HTA contexts been taken account. when developing or transforming the method?<br>✓ Go to R2.3.2 & A2.3.3. | R2.3.2<br>✓ If Yes<br>Report techniques used to evaluate transferability (e.g. interviews with practitioners); List all HTA stakeholders who evaluate transferability, and report contexts they belong to (e.g. geographic and therapeutic areas).<br>✓ If No<br>Discuss how it could impact transferability of the method. |
| A2.3.3 Can the HTA method be disseminated and implemented in various contexts, considering the validity, applicability, and transferability of the method prototype and its derivative versions?<br>✓ If Yes, go to R2.3.3 & A3.1.1.<br>✓ If No, go to R2.3.3 & A2.1.1, for another round of method development. | R2.3.3<br>✓ If Yes or No<br>Report the process on how the decision on dissemination and implementation is made; Describe contexts where an HTA method is disseminated or implemented; Specify potential causes of preventing dissemination and implementation (e.g. design flaws, lack of transferability, or wrong operations). |
| ***Implementation Phase - 3.1 - Plan for Implementation*** | |
| A3.1.1 Has the HTA method been disseminated or diffused?<br>✓ Go to R3.1.1 & A3.1.2. | R3.1.1<br>✓ If Yes<br>Report approaches used for dissemination (e.g. training practitioners) or diffusion (e.g. scientific publications); Report the involvement of developers, practitioners, and beneficiaries in the action of dissemination.<br>✓ If No<br>Specify the reason. |
| ***Implementation Phase - 3.1 - Plan for Implementation*** | |

**Appendix 2. Continued**

| Actions for innovating an HTA method | Reporting a process of innovation |
|---|---|
| A3.1.2 Has an implementation strategy been developed, for guiding the resources needed for conducting and monitoring the implementation, and for motivating potential practitioners to adopt the novel HTA method?<br>✓ Go to R3.1.2 & A3.2.1. | R3.1.2<br>✓ If Yes<br>Report the implementation strategy; Report resource needed for conducting and monitoring the implementation; Report planned approaches used to motivate practitioners in real-world practice.<br>✓ If No<br>Describe the impact of method implementation without an implementation strategy. |
| ***Implementation Phase - 3.2 - Apply a Method To Practice*** | |
| A3.2.1 Is technical assistance available from developers during method implementation in real-world practice?<br>✓ Go to R3.2.1 & A3.2.2. | R3.2.1<br>✓ If Yes<br>Describe the ways developers can be approached; Describe what type of technical assistance developers can provide.<br>✓ If No<br>Specify the reason, and evaluate the impact |
| A3.2.2 Is the method implementation continuously monitored by developers, and is feedback from practitioners and beneficiaries accessible to developers?<br>✓ If Yes, go to R3.2.2 & A3.2.3.<br>✓ If No, go to A3.2.3. | R3.2.2<br>✓ If Yes<br>Describe how method implementation is monitored in practice; Report the technique(s) used to obtain feedback; Summarize feedback.<br>✓ If No<br>Specify the reason. |
| A3.2.3 Have implementation strategies been adjusted during implementation?<br>✓ Go to R3.2.3 & A3.3.1. | R3.2.3<br>✓ If Yes<br>If available, report reasons why implementation strategies are adjusted (e.g. tailored contexts); Describe the adjusted implementation strategies; Describe impact of the adjusted implementation strategies.<br>✓ If No<br>Specify the reason. |
| ***Implementation Phase - 3.3 - Test & Transfer*** | |
| A3.3.1 Have information on validity of an HTA method been obtained from various contexts during method implementation?<br>✓ Go to R3.3.1 & A3.3.2. | R3.3.1<br>✓ If Yes<br>Report validity of the method; Report contexts where information on method validity are obtained.<br>✓ If No<br>Specify the reason. |
| A3.3.2 Have information on adoption of an HTA method been obtained from various contexts, during method implementation?<br>✓ Go to R3.3.2 & A3.3.3. | R3.3.2<br>✓ If Yes<br>Report the extent to which practitioners and beneficiaries adopt the method; Report contexts where relevant information are obtained.<br>✓ If No<br>Specify the reason. |

**Appendix 2. Continued**

| Actions for innovating an HTA method | Reporting a process of innovation |
|---|---|
| ***Implementation Phase - 3.3 - Test & Transfer*** | |
| A3.3.3 According to R1.3.3 & R3.3.1 & R3.3.2, could the method be directly applied to all contexts of interest?<br>✓ If Yes, go to R3.3.3 & A1.1.1, for another round of identifying limitations of existing HTA methods.<br>✓ If No, go to R3.3.3 & A2.1.1. | R3.3.3<br>✓ If Yes<br>Specify the reason.<br>✓ If No<br>Describe contexts where a method could be directly applied; Describe contexts where a method could not be directly applied; Specify facilitators and barriers of transferring the method. |

# Part 2

## Quality Assessment of Studies Using RWD for HTA

Chapter 4

# Methodological Quality of Retrospective Observational Studies Investigating Effects of Diabetes Monitoring Systems: a Systematic Review

Li Jiu, Junfeng Wang, Maria Kamusheva, Maria Dimitrova, Konstantin Tachkov, Petya Milushewa, Zornitsa Mitkova, Guenka Petrova, Rick Vreman, Aukje K Mantel-Teeuwisse, Wim G Goettsch

*Submitted*

# Abstract

## Background

Retrospective observational studies (ROSs) have been frequently used to investigate treatment effects of diabetes monitoring systems (DMS), i.e. medical devices to monitor blood glucose. However, due to quality concerns, the findings of such studies were often questioned by clinical, regulatory, or health technology assessment decision-makers. We aimed to conduct a systematic review to assess the methodological quality of ROSs investigating DMS effects, and to explore the trend in quality change over time.

## Methods

Embase, PubMed, Web of Science, and Scopus were systematically searched for English-language articles published from January 2012 to March 2021. Randomized controlled trials or other prospective studies were manually excluded. The ROBINS-I (Risk Of Bias In Non-randomized Studies – of Interventions) was used for assessing RoB. To investigate the quality change over time, we divided the study into three subgroups according to publication year, and compared the proportion of studies with the same quality level among the three subgroups.

## Results

We identified 4926 articles, of which 72 were eligible for inclusion. Twenty-six studies were published before 2018, 22 in 2018 or 2019, and 24 after 2019. The overall methodological quality was quite low, as 61 (85%) studies were graded as facing critical or serious RoB. Also, the overall methodological quality did not substantially improve over time. The major contributors to low quality included confounding, missing data, and selection of the reported results.

## Conclusions

The retrospective observational studies investigating DMS effects generally had a high risk of bias, and this did not substantially improve in the past ten years. Thus, clinical, regulatory, or HTA decision-makers may need strategies to effectively exploit these suboptimal studies. Also, to further improve study quality, extra efforts may be needed, such as guiding the tool selection regarding quality improvement in the tools.

## Introduction

A retrospective observational study (ROS) is a non-randomized and non-interventional study of existing data, such as electronic medical records or patient registries, to measure outcomes of interest (1,2). Compared to prospective studies, including clinical trials, ROSs are generally more financially feasible and time-saving (3), and enable outcome measurement from a larger sample population (4). Hence, ROSs are especially useful when investigating chronic diseases, where long-term follow-ups and large sample sizes are needed.



Given the strengths of retrospective observational studies and the growing trends of utilizing routinely collected data for evaluating medical devices (5), retrospective observational studies have been popular in investigating outcomes of diabetes monitoring systems (DMSs). A DMS, also called a glucose monitoring system, is a type of portable medical device for monitoring blood glucose in diabetic patients (6,7). Based on device characteristics (e.g. how blood glucose is measured, how an outcome event is alarmed, and how an insulin therapy is adjusted), DMSs can be classified into multiple types, including but not limited to continuous glucose monitoring (CGM), self-monitoring of blood glucose (SMBG), closed-loop system, and sensor-augmented pump therapy (SAPT) (8).CGM can be further classified into several variants, including professional CGM (e.g. retrospective CGM) and personal CGM (e.g. flash CGM and real-time CGM) (9,10).

Although ROSs have dramatically increased in numbers, their findings have not been fully trusted by clinical, regulatory, or health technology assessment (HTA) decision-makers. According to the hierarchy of evidence proposed repeatedly in the past 20 years, which was valued by all decision-makers (11-13), ROSs are usually assigned a lower grade than prospective studies with the same design (e.g. retrospective cohort vs. prospective cohort) (12-14). The lack of trust was also partially confirmed by the lack of synthesis of ROSs in meta-analyses, which are often considered a gold standard of evidence by decision-makers (15-17). According to a review, 16 meta-analyses investigating DMS outcomes were published before 2020, but only two included ROSs as evidence (8).

One plausible explanation for the untrustworthiness is that ROSs suffer from biases more frequently than prospective studies (18). For example, as the data from databases are originally collected for a different purpose, they are hardly controlled by researchers, and might be incomplete, inaccurate, or inconsistently measured (18,19). This would consequently increase the risk of information bias (19).

Also, recall bias is an major issue for retrospective studies, because self-reporting of outcomes may be needed, and disease status may influence the ability to accurately recall prior exposures (20).

To improve quality of ROSs and to establish credibility of their findings, many efforts have been taken in the past ten years, such as developing and applying advanced methods for bias adjustment and developing tools for researchers (21,22). However, we don't know whether the existing efforts have contributed to the improved quality, in terms of risk of bias (RoB). Hence, our study aims to systematically identify retrospective observational studies investigating effects of diabetes monitoring systems, to evaluate their RoB, and to explore the trend in quality over time. This research was performed as part of the HTx project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825162 (23).

# Methods

## Protocol registration

To conduct a systematic review, we first registered a study protocol (CRD42021273217) in the PROSPERO (International Prospective Register of Systematic Reviews) database. To ensure the transparency, completeness, and accuracy of the review, we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) 2020 statement (24).

## Search strategy

A systematic literature search with three key concepts (i.e. diabetes, DMS, and glycemic outcomes) was conducted to identify ROSs for investigating outcomes of diabetes monitoring systems, which were published between 1st September, 2011 and 31st, March 2021. We used the concept "glycemic", as it could cover all major clinical DMS outcomes (6). Also, to avoid exclusion of eligible studies not using the direct wording "retrospective", we did not add it in the search string. We searched PubMed, Embase, Web of Science, and Scopus, and screened the reference lists of studies considered eligible in the full-text review. The main database search was conducted on 1st April, 2021. The search strategy was developed by two authors (LJ & JW), then edited by an experienced librarian in document retrieval from Utrecht University. The detailed search strategy is shown in Appendix 1.

## Eligibility criteria

A study was included if (1) the target population were patients diagnosed with diabetes regardless of type of diabetes and age group; (2) either the intervention or comparator(s) was a diabetes monitoring system; (3) primary outcomes included at least one glycemic outcome, including hemoglobin A1c (HbA1c), time in range of blood glucose, glycemic variability, hypoglycemia, and hyperglycemia, etc.; (4) data were extracted from databases, including (electronic) medical (or health) records, patient registries, healthcare administrative data(i.e. administrative claims), and patient reported outcomes, etc.; (5) the study was published in English. A study was excluded if it was a clinical trial (either randomized or nonrandomized) or prospective cohort study.



## Study identification

One author (LJ) screened all titles and abstracts of identified records, while two other authors (JW or RV) each independently screened a random set of 5%. At a later stage, the full-texts were reviewed independently by two authors (LJ and MK). Any discrepancy during study identification was solved through discussion between two authors.

## Data extraction

A form to extract study characteristics was developed by LJ and then adjusted by JW, RV, WG, and AMT. The data items included target populations (i.e. age groups (e.g. adults), types of diabetes, regions, and primary or secondary settings), types of DMSs for interventions or comparators, diabetes medication, outcomes of interest, study designs (e.g. cohort studies, case control studies, etc.), data sources, sample sizes, lengths of follow-ups, and publication years. Study characteristics were collected by one author (LJ).

## Quality assessment

A quality assessment form was developed based on the ROBINS-I (Risk Of Bias In Non-randomized Studies – of Interventions)25 tool. The reason for using ROBINS-I was that it provided a comprehensive list of signaling questions used for assessing RoB in seven domains of non-randomized studies, and that it was proved robust and highly recommended by authorities, e.g. Cochrane for quality assessment in systematic reviews (26).

For quality assessment, all eligible studies were randomly divided into five parts, then each part was independently assessed by two authors (LJ, MK, KT, MD, GP, or PM). To guarantee the reliability, one author who was experienced with the two appraisal tools (LJ) participated in quality assessment and discrepancy discussion of all studies and shared the interpretation of signaling questions from the previous pair of authors with the next. Any disagreement in quality assessment was discussed and resolved by two authors.

### Data analysis

The characteristics of the selected studies were presented as numbers and percentages. For quality assessment, we estimated the proportion of studies that fulfilled each signaling question. Also, we rated the overall quality as "low RoB", "moderate RoB", "serious RoB", "critical RoB", or "no information", based on algorithms provided by ROBINS-I. To investigate the methodological quality over time, we first divided the studies into three subgroups based on publication years, and ensured that the number of studies in each subgroup was approximately similar. Then we compared the proportion of studies with the same overall or domain-specific quality level (e.g. moderate RoB) among the three subgroups.

## Results

### Selection of studies for quality assessment

As shown in Figure 1, We identified 4926 records after removing duplicates, and excluded 4799 records after reviewing titles and abstracts. Of the remaining 127 records, 72 were finally included for the quality assessment after the full-text review. References of the included studies are shown in Appendix 2. Among the 72 studies, 26 were published before 2018, 22 in 2018 or 2019, and 24 after 2019.

### Study characteristics

Characteristics of the included studies are presented in Table 1 and Appendix 3. More than two-thirds of the studies focused on a single type of diabetes, either type 1 (55%) or type 2 (22%), while the rest focused on multiple types. In addition, the age groups of the population varied significantly among the studies. Regarding DMSs investigated, CGM was the most frequently used intervention (69%), while SMBG and SAPT ranked second (14%) and third (8%), respectively. Most studies included HbA1c as outcome, while hypoglycemia, hyperglycemia, and time in range of blood glucose ranked second to fourth. Almost half of the studies utilized the cohort study design, while the other half were crossover, cross-sectional, case-control, or panel studies. More than four-fifths of the studies used electronic health or medical records. In terms of sample size, more than one-third of the studies had a small sample size, which was less than 100 patients.

**Table 1. Study characteristics of the included 72 studies**

| | N (total number = 72) | % |
|---|---|---|
| **Population** | | |
| ***Disease type*** | | |
| Type 1 diabetes only | 40 | 55% |
| Type 2 diabetes only | 16 | 22% |
| Type 1 and type 2 diabetes | 10 | 14% |
| Other[a] | 6 | 8% |
| ***Age group***[b] | | |
| Adult | 36 | 49% |
| Non-Adult | 8 | 11% |
| All | 19 | 26% |
| Other | 9 | 12% |
| ***Region*** | | |
| Asia | 17 | 23% |
| North America | 25 | 34% |
| Europe | 25 | 34% |
| Africa | 1 | 1% |
| Oceania | 1 | 1% |
| Transcontinental | 3 | 4% |
| **Population** | | |
| ***Setting*** | | |
| Primary care | 12 | 16% |
| Secondary care | 33 | 45% |
| Primary & Secondary care | 5 | 7% |
| Not reported | 23 | 32% |
| **Diabetes monitoring system (DMS)** | | |
| ***Intervention*** | | |
| Continuous glucose monitoring (CGM)[c] | 50 | 69% |
| Self-monitoring of blood glucose (SMBG) | 10 | 14% |
| Sensor-augmented pump therapy (SAPT) | 6 | 8% |
| Closed-loop system (artificial pancreas) | 4 | 5% |
| Other[d] | 3 | 4% |
| ***Comparator*** | | |
| Continuous glucose monitoring (CGM) | 3 | 4% |
| Self-monitoring of blood glucose (SMBG) | 5 | 7% |

**Table 1. Continued**

| | N (total number = 72) | % |
|---|---|---|
| Sensor-augmented pump therapy (SAPT) | 4 | 5% |
| Non-user[e] | 26 | 36% |
| No comparator[f] | 34 | 47% |
| **Diabetes medication[g]** | | |
| Insulin therapy | 29 | 40% |
| Non-insulin therapy | 6 | 8% |
| Insulin & Non-insulin therapy | 19 | 26% |
| No reported | 18 | 25% |
| **Outcome (not mutually exclusive)** | | |
| hemoglobin A1c (HbA1c) | 63 | 86% |
| Time in range of blood glucose | 11 | 15% |
| Fasting blood glucose | 5 | 7% |
| Hypoglycemia | 19 | 26% |
| Hyperglycemia | 11 | 15% |
| Other | 36 | 49% |
| **Study design** | | |
| ***Costs & effectiveness*** | | |
| Effectiveness | 28 | 38% |
| Comparative effectiveness | 37 | 51% |
| Cost-effectiveness | 1 | 1% |
| (Comparative) Effectiveness & Costs[h] | 6 | 8% |
| ***Study type*** | | |
| Cohort | 35 | 48% |
| Crossover | 17 | 23% |
| Cross-sectional | 11 | 15% |
| Case-control | 8 | 11% |
| Panel | 1 | 1% |
| **Data source** | | |
| ***Single data sources (n=67)*** | | |
| Electronic medical records | 56 | 77% |
| Healthcare administrative data | 5 | 7% |
| Registry | 4 | 5% |
| Audit data | 2 | 3% |
| ***Multiple data source (n=5)*** | | |
| Electronic medical records & Patient-reported outcomes | 3 | 4% |

**Table 1. Continued**

| | N (total number = 72) | % |
|---|---|---|
| Electronic medical records & Healthcare administrative data | 1 | 1% |
| Registry & Patient-reported outcomes | 1 | 1% |
| **Sample size** | | |
| ***Intervention (n=72)*** | | |
| >1000 | 14 | 19% |
| 100-1000 | 27 | 37% |
| <100 | 31 | 42% |
| ***Comparator (n=39)*** | | |
| >1000 | 8 | 21% |
| 100-1000 | 16 | 41% |
| <100 | 15 | 38% |
| **Length of follow-up** | | |
| No follow-up[i] | 9 | 12% |
| <=3 months | 49 | 67% |
| >3 months | 3 | 4% |
| Not reported | 12 | 16% |

[a] The study did specify the type of diabetes, such as Type-1, Type-2, gestational diabetes.

[b] "Other" indicates age groups which included both adults and non-adults, rather than all ages, e.g., from year 2 to year 72.

[c] "Continuous glucose monitoring (CGM)" includes all its subsets, including flash glucose monitoring (FGM), professional CGM, real-time CGM (rtCGM), retrospective CGM, intermittent scanning CGM (iscCGM).

[d] "Other" indicates nurse-directed Electronic Glycemic Management System (eGMS) or self-monitoring of blood glucose combined with telemedicine.

[e] "Non-user" indicates the comparator group of the study did not use a DMS. For example, an intervention group used CGM while the comparator group only received an insulin and/or a non-insulin therapy without CGM.

[f] "No comparator" indicates the study did not set a comparator group.

[g] "Diabetes medication" indicates the medication patients received in addition to the DMS. It includes insulin therapy, such as insulin pump, multiple daily insulin injection (DMI), and continuous subcutaneous insulin infusion (CSII), and non-insulin therapy, such as oral hypoglycemic agent (OHA), medication (e.g. metformin, metformin combination, sulfonylurea, dipeptidyl peptidase-4 inhibitor, thiazolidione), and healthy diet.

[h] "Effectiveness & Costs" indicates the study which included both effectiveness and costs, rather than cost-effectiveness, as outcomes,.

[i] "No follow-up" refers to cross-sectional studies.

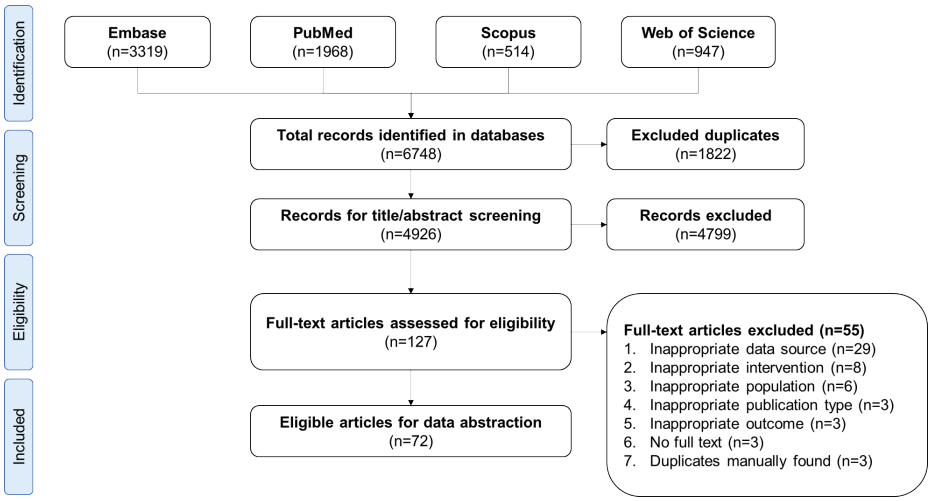



**Figure 1. The flow chart for the inclusion and exclusion of studies.**

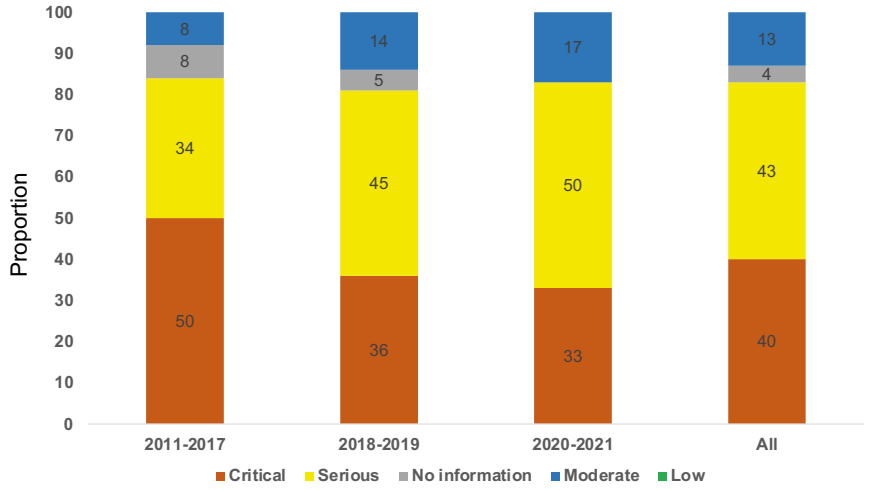
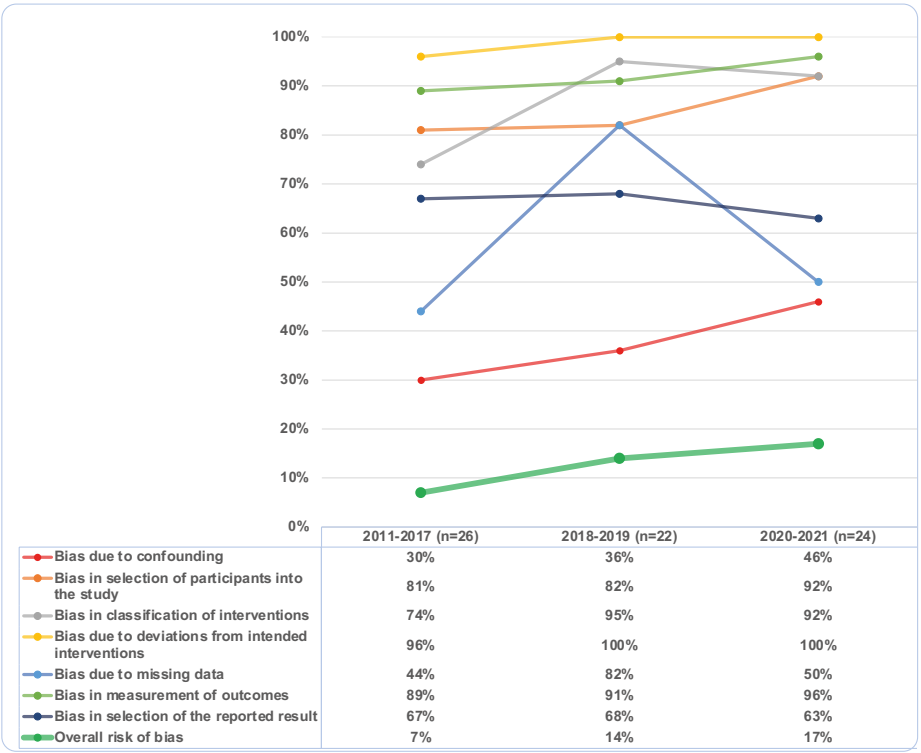



**Figure 2. Change of overall quality of studies over time (ROBINS-I) (n=72).**




| | 2011-2017 (n=26) | 2018-2019 (n=22) | 2020-2021 (n=24) |
|---|---|---|---|
| Bias due to confounding | 30% | 36% | 46% |
| Bias in selection of participants into the study | 81% | 82% | 92% |
| Bias in classification of interventions | 74% | 95% | 92% |
| Bias due to deviations from intended interventions | 96% | 100% | 100% |
| Bias due to missing data | 44% | 82% | 50% |
| Bias in measurement of outcomes | 89% | 91% | 96% |
| Bias in selection of the reported result | 67% | 68% | 63% |
| Overall risk of bias | 7% | 14% | 17% |

**Figure 3. Proportion of studies with low or moderate RoB in the three time periods for different RoB domains according to ROBINS-I (n=72).**

## Quality assessment

The overall methodological quality of studies, as well as the change of quality over time, are shown in Figure 2. In summary, the overall methodological quality was poor. As shown in Figure 2, 29 (40%), 31 (43%), 9 (13%), and 3 (4%) studies were graded as facing critical RoB, serious RoB, moderate RoB, and having no information, respectively. No studies were graded as having low RoB. Also, the studies published in the three time periods (i.e. 2011-2017, 2018-2019, 2020-2021) differed slightly and showed a slight growing trend over time, in terms of methodological quality. More specifically, the proportion of studies with critical overall RoB decreased from 50% to 36%, then to 33%, while the proportion of studies with moderate RoB increased from 8% to 14% then to 17%.

In contrast, the trend of quality in the seven RoB domains differed in patterns. As shown in Figure 3, the proportion of studies with low or moderate RoB experienced a large increase (i.e. >15%) only in two domains: bias due to confounding and bias in classification of interventions. However, regarding bias related to participant

selection, deviations from intended interventions, and outcome measurement, the proportion increased only slightly (i.e. <=8%) over time, or even remained the same. Also, no obvious change in quality was detected regarding bias in selection of the reported result. Unlike the above-mentioned domains, studies published in 2018 or 2019 had significantly lower RoB due to missing data than those published before or after this time period.

In general, the proportion of studies with low or moderate overall RoB was much lower than the proportions regarding domain-specific RoB. This is reasonable, as according to the ROBINS-I, a study was graded as low or moderate RoB only if all the seven domains were graded as low or moderate RoB. Furthermore, Figure 3 shows that the studies published in the three time periods shared similar ROB contributors, i.e. bias related to confounding, missing data, and selection of the reported result.

The details on how the ROSs fulfilled each signaling question within the seven ROBINS-I domains are shown in Figure 4 and Appendix 4. The number one contributor to low overall methodological quality was confounding bias, which was caused by confounders (i.e. variables distorting associations between an intervention and outcome25). In this domain, 45 (62%) of all studies were graded as critical or serious RoB. More specifically, of 55 (76%) studies for which assessing time-varying confounding was not necessary, only 32 used an appropriate method (e.g. regression, stratification, and matching) to control for confounders (signaling question 1.4), and 21 provided clear evidence that confounders were measured validly and reliably (signaling question 1.5). Similarly, of the 17 (24%) studies which needed assessment of time-varying confounding, (i.e. with an answer "Yes", "Probably Yes", or "No information" in signaling question 1.3) 11 provided adjustment techniques, and only 8 measured confounders validly and reliably (signaling question 1.7, 1.8). The confounding domain also explained why no studies were graded as low overall RoB. According to the ROBINS-I, only studies that were equivalent to an RCT can be graded as low overall RoB, and the included ROSs did not meet the criterion.

The number two contributor to low overall methodological quality was bias due to missing data, as 25 (34%) studies were graded as having this serious domain-specific RoB. In 8 studies, outcome data were not available for some patients (signaling question 5.1). Also, about 40% of all studies excluded patients due to missing data on intervention status or other variables needed for analyses (e.g. confounders), or did not specify anything on missing data (signaling question 5.2, 5.3). Of the studies with concerns on missing data, only 17 provided clear evidence that missing

data would not bias effect estimates (signaling question 5.4), and only 11 improved robustness of results by using appropriate methods to handle missing data or by performing sensitivity analyses to assess underlying assumptions about missing data (signaling question 5.5).

Similarly, concerning selection of reported results, 24 (33%) studies were graded as facing critical or serious RoB, for several reasons. First, 17 (23%) studies failed to provide estimates for all patient subgroups derived from the whole database(s) (signaling question 7.3). Second, in 14 (19%) studies, outcomes were measured with multiple techniques, but only one effect estimate was reported (signaling question 7.1). Lastly, in 5 (7%) studies, data analysis was performed with more than one method, but only one effect estimate was reported (signaling question 7.2).



## Discussion

We conducted a systematic review of retrospective observational studies which investigated effects of diabetes monitoring systems to evaluate and compare quality of the identified 72 studies that were published in three time periods (i.e. 2011-2017, 2018-2019, 2020-2021). In general, the overall quality of retrospective observational studies was poor, and the quality only slightly improved over time. The main contributor of low quality in our review included RoB due to confounding, missing data, and selection of the reported results.

Our finding regarding overall RoB was consistent with previous reviews which used the ROBINS-I to evaluate quality of retrospective observational studies in the field of diabetes. Kumar et al. (2022) identified 11 retrospective cohort studies which investigated effects of CGM and SMBG in the management of cystic fibrosis related diabetes (27).In their review, seven studies were rated as having critical or serious RoB, while no study was rated as having low RoB. Similarly, Islam et al. (2022) identified 11 retrospective studies investigating sulfonylureas in diabetic patients and rated nine of them with overall critical or serious RoB (28).Golden et al. (2012) did not use the ROBINS-I but the Downs and Black quality checklist supplemented with items from the Methods Guide for Effectiveness and Comparative Effectiveness Reviews (29). Still, they confirmed our finding by showing that none of the identified retrospective observational studies for investigating CGM and SMBG was rated as good quality (29). However, our findings regarding domain-specific RoB were not completely consistent with previous reviews, especially regarding bias due to missing data. We found missing data as the major RoB contributor, while Kumar et al. (2022) and Islam et al. (2022) graded


s there potential for confounding of the effect of intervention in this study?
72
Was the analysis based on splitting participants' follow up time according to intervention received?
21
51
Were intervention discontinuations or switches likely to be related to factors that are prognostic for the outcome?
8
9
4
)id the authors use an appropriate analysis method that controlled for all the important confounding domains?
32
3
20
Were confounding domains that were controlled for measured validly and reliably by the variables available in this study?
21
7
7
)id the authors control for any post-intervention variables that could have been affected by the intervention?
16
35
)id the authors use an appropriate analysis method that adjusted for all the important confounding domains and for time varying confounding?
10
7
Were confounding domains that were adjusted for measured validly and reliably by the variables available in this study?
8
2
Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention?
7
1
64
Were the postintervention variables that influenced selection likely to be associated with intervention?
6
2
Were the postintervention variables that influenced selection likely to be influenced by the outcome or a cause of the outcome?
1
)o start of follow-up and start of intervention coincide for most participants?
48
10
14
Were adjustment techniques used that are likely to correct for the presence of selection biases?
18
10
Were intervention groups clearly defined?
66
6
Was the information used to define intervention groups recorded at the start of the intervention?
71
1
Could classification of intervention status have been affected (blinded) by knowledge of the outcome or risk of the outcome?
4
64
Were there deviations from the intended intervention beyond what would be expected in usual practice
72
Were these deviations from intended intervention unbalanced between groups and likely to have affected the outcome?
Were outcome data available for all, or nearly all, participants?
64
8
Were participants excluded due to missing data on intervention status?
27
5
40
Were participants excluded due to missing data on other variables needed for the analysis?
26
8
38
re the proportion of participants and reasons for missing data similar across interventions?
17
17
s there evidence that results were robust to the presence of missing data?
11
7
16
Could the outcome measure have been influenced by knowledge of the intervention received?
13
2
57
Were outcome assessors aware of the intervention received by study participants?
71
1
Were the methods of outcome assessment comparable across intervention groups?
67
3
2
Were any systematic errors in measurement of the outcome related to intervention received?
14
12
46
s the reported effect estimate likely to be selected, on the basis of the results, from multiple outcome measurements within the outcome domain?
14
5
53
s the reported effect estimate likely to be selected, on the basis of the results, from multiple analyses of the intervention-outcome relationship?
4
64
s the reported effect estimate likely to be selected, on the basis of the results, from different subgroups?
17
7
48
Y/PY
NI
N/PI
5
10
15
20
25
30
35
40
45
50
55
60
65
70
Number of studies (N=72)

< **Figure 4. Number of studies that fulfilled each signaling question (ROBINS-I) (n=72).**
The signaling questions, i.e. 1.3-1.8, 2.2, 2.3, 2.5,4,5.5, are only applicable to a part of studies, according to algorithms defined by the ROBINS-I.

more than half of ROSs in this domain as low or moderate RoB (27,28). One explanation for the inconsistency may be that we identified a larger proportion (almost 50%) of studies which clearly indicated that patients were excluded due to missing data, and that only about 30% of these studies adjusted for potential bias, e.g. using multiple imputation. We also included a larger number of studies (72 vs.11). For the other RoB domains, our findings and those from previous reviews were relatively more consistent. For example, Islam et al. (2022) confirmed that bias due to confounding and selection of the reported result were major contributors, as ten and seven of the 11 ROSs were graded as critical or serious RoB, respectively (28).



Our findings demonstrate the barriers of applying retrospective observational studies to clinical, regulatory, or HTA decision-making (30-32). However, given the benefits of ROSs, especially when conducting a prospective study was not feasible or too expensive (3,21), collaborative efforts may be needed to properly apply these studies with suboptimal quality. In addition, to following general recommendations proposed by previous research, such as filling evidence gap when RCTs are not available (32,33), we recommend decision-makers to improve the alignment of decision-making criteria, and to put more weight on major contributors to low quality. Our review shows that, with regard to DMSs, confounding, missing data, and selective reporting need special attention from decision-makers. Future studies may be needed to continuously monitor the methodological quality, and to check whether improvements are found.

Our review also implied that, given the only slightly increasing trends, the existing efforts to improving quality of ROSs might help, but could not fully address relevant RoB concerns. One type of efforts is to disseminate robust statistical methods to researchers for correcting the lack of randomization in ROSs. For example, Oliwier et al. (2020) has recommended the use of propensity score matching, structural modeling with graph-analytic approaches, and component analysis (21). By adopting these techniques, the lack of randomization in ROSs could be corrected, and the induced bias, such as confounding or selection bias, could be mitigated. To further promote these techniques, a few tools have been specifically designed for researchers, with an aim to support the development of scientifically rigorous observational research. Some of the tools included the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Checklist, the CER-Collaborative (Comparative Effectiveness Research Collaborative: Observational Study Assessment Questionnaire) checklist, and the Patient-Centered Outcomes Research Institute Methodology

Standards (PCORI) (22). However, according to our findings, these tools might not be commonly adopted by researchers in the field of diabetes, even though they have been repeatedly promoted by statisticians and tool developers.

One explanation for lack of adoption was that, ROS researchers became overwhelmed by a complex set of tools, which differed significantly in number, content, or format of signaling questions (22,34). As emphasized by Buccheri et al. (2017), the failure of selecting an appropriate tool could make quality assessment less effective and more laborious (35). Similarly, if more than one tool could potentially satisfy the researchers' needs, the variety of tools could make the tool combination extremely complicated. For example, the CER-Collaborative could also be potentially applied for ROSs, as it provides recommendations which are tailored to ROSs (36). This tool includes six domains corresponding to six phases in a process of conducting a non-randomized study (e.g. study design, data collection, data analysis, and reporting, etc.), and each domain is related to multiple biases. In opposite, the ROBINS-I includes seven ROB domains, and each domain corresponds to multiple phases of the process. Though some signaling questions from the two tools are similar in content (e.g. confounding), they differ in how they are raised.

Even if ROS researchers have successfully identified one or several robust tools, they may not be well-informed on what these tool(s) cannot tell. The reason is that the tools might not mention the disputes on RoB adjustment which they could not resolve. For example, in our review, approximately 40% of studies had a sample size of less than 100, and we found no consensus among the RoB tools on how the sample size should be addressed or discussed. The lack of consensus was also reflected in recent debates. For example, there were opposite opinions on what method (e.g. post-hoc power analysis, Bayesian analyses, etc.) should be used for calculating statistical power of a ROS, and whether a method was valid (37-39). There was also debate on whether a small sample size was a real problem of study reliability that needed to be addressed (39-41). While these debates existed, they were rarely reflected in existing tools. Even the tools which provided tailored recommendations for ROSs, such as the CER-Collaborative or REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement (42), they merely admitted that sample size was an issue in ROSs, but did not provide a solution. Hence, we suggest the tools should not only state what they can provide, but also, if available, the debates that they cannot address. In this way, ROS researchers could attach more importance to potential problems that reduces study quality, and tool developers could understand where the existing tools can be further improved.

In addition, the almost unchanged high RoB over time may be explained by the relatively inefficient ways of disseminating methods for addressing RoB. Though relevant tools were published (22,34), channels were lacking for providing detailed illustration of the tools. According to Whiting et al. (2017), strategies beyond publications were recommended to promote dissemination of tools, including designing a tool website, translating a tool into other languages, and encouraging uptake of tools by leading organizations (43). Hence, we recommend further research to explore how existing tools were disseminated and the association between dissemination strategies and user adoption. To encourage the researchers' adoption, we also recommend to develop and implement a website or app, which interactively guides the selection and combination of appraisal tools based on the user needs (e.g. evaluating quality of ROSs). Such a website has already been developed for human observational studies investigating exposures (44),but it is not available for studies investigating interventions. Furthermore, we suggest that, the journal which accepts ROSs should be more consistent and clear on which tools they would need before they accept a ROS.



## Limitations

Our study has some limitations. One limitation is that only one author scanned all titles and abstracts, and only 10% were independently scanned by another author. Even, as shown in our results, no additional eligible study was identified by the second author, some potentially eligible studies might be missing. However, this limitation probably will not bias the assessment and comparison of study quality, because the cause of missed studies was unrelated to study characteristics or publication years. In future, advanced methods to facilitate selecting more timely and reliable reviews may be applied. Text-mining can be promising, because it could help reduce number of hits that need to be screened manually, and could act as a second author for study identification (45,46). To refine and apply these text-mining methods, reviews of diabetes monitoring systems may act as a case study. Another limitation is that all studies were appraised independently by two authors, but the different pairs of authors might interpretate a signaling question differently. We mitigated this potential limitation through involving one author with experience with appraisal tools in quality assessment and discrepancy discussion of all studies.

In addition, our review only included reviews published in English, and the studies only investigating non-glycemic outcomes were excluded. We did not expect these would alter our findings, because language and the selection of target outcomes were unlikely to have a causal inference to study quality. However, researchers who investigate patient adherence to diabetes monitoring systems or compare DMS performance in non-English-speaking regions should take this limitation into account.

## Conclusions

The retrospective observational studies investigating DMS effects generally had a high risk of bias, and this did not substantially improve over time. Thus, clinical, regulatory, or HTA decision-makers may need strategies to effectively exploit these suboptimal studies. Also, strategies to help researchers improve study quality are needed, such as guiding the tool selection.

## Author contribution

LJ edited the search strategy, scanned all titles and abstracts, reviewed all full-text articles, extracted study characteristics, assessed quality of identified studies, solved discrepancies regarding quality assessment, analyzed the data, and wrote the manuscript. All co-authors contributed to this study and critically reviewed and approved the manuscript.

# References


1. Gardner AL, Charlesworth M. How to write a retrospective observational study. Anaesthesia. 2022.
2. Hess DR. Retrospective studies and chart reviews. Respir Care. 2004;49(10):1171-4.
3. Talari K, Goyal M. Retrospective studies–utility and caveats. J R Coll Physicians Edinb. 2020;50(4):398-402.
4. Anthonisen NR. Retrospective studies. Can Respir J. 2009;16(4):117-8.
5. Ciani O, Federici C, Tarricone R. Current and future trends in the HTA of medical devices. InXIV Mediterranean Conference on Medical and Biological Engineering and Computing. 2016; 1345-1348.
6. Whitmore C. Blood glucose monitoring: an overview. Br J Nurs. 2012;21(10):583-7.
7. Bolla AS, Priefer R. Blood glucose monitoring-an overview of current and future non-invasive devices. Diabetes Metab Syndr. 2020;14(5):739-51.
8. Kamusheva M, Tachkov K, Dimitrova M, Mitkova Z, García-Sáez G, Hernando ME, et al. A systematic review of collective evidences investigating the effect of diabetes monitoring systems and their application in health care. Front Endocrinol. 2021;12:636959.
9. Mancini G, Berioli MG, Santi E, Rogari F, Toni G, Tascini G, et al. Flash glucose monitoring: a review of the literature with a special focus on type 1 diabetes. Nutrients. 2018;10(8):992.
10. Chamberlain JJ. Continuous glucose monitoring systems: categories and features. 1st ed. Arlington (VA): American Diabetes Association; 2018.
11. Velasco-Garrido M, Busse R. Assessing research. In: Health technology assessment: an introduction to objectives, role of evidence, and structure in Europe. World Health Organization. 2005.
12. Petrisor BA, Bhandari M. The hierarchy of evidence: levels and grades of recommendation. Indian J Orthop. 2007;41(1):11.
13. Schlegl E, Ducournau P, Ruof J. Different weights of the evidence-based medicine triad in regulatory, health technology assessment, and clinical decision making. Pharmaceut Med. 2017;31(4):213-6.
14. Manterola C, Asenjo-Lobos C, Otzen T. Hierarchy of evidence: levels of evidence and grades of recommendation from current use. Rev Chilena Infectol. 2014;31(6):705-18.
15. Chaiyakunapruk N, Saokaew S, Sruamsiri R, Dilokthornsakul P. Systematic review and network meta-analysis in health technology assessment. J Med Assoc Thai. 2014;97:S33-42.
16. Ren S, Oakley JE, Stevens JW. Evidence synthesis for health technology assessment with limited studies. Value Health. 2017;20(9):A770.
17. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg. 2011;128(1):305.
18. Bashir MM, Maskari FA, Ahmed L, Al-Rifai RH. Prospective Vs Retrospective Cohort Studies: Is a Consensus Needed?. Int J Epidemiol. 2021; 50(Supplement_1):dyab168-063.
19. Ramirez-Santana M. Limitations and biases in cohort studies. In: RM Barría, editor. Cohort Studies in Health Sciences. 1st ed. London: Intechopen ;2018.
20. Song JW, Chung KC. Observational studies: cohort and case-control studies. Plast Reconstr Surg. 2010; 126(6): 2234.
21. Dziadkowiec O, Durbin J, Jayaraman Muralidharan V, Novak M, Cornett B. Improving the quality and design of retrospective clinical outcome studies that utilize electronic health records. HCA Healthc J Med. 2020;1(3):4.
22. Morton SC, Costlow MR, Graff JS, Dubois RW. Standards and guidelines for observational studies: quality is in the eye of the beholder. J Clin Epidemiol. 2016;71:3-10.

23. HTx: About HTx project. Available from: https://www.htx-h2020.eu/about-htx-project. [Accessed Oct 25, 2022].
24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev. 2021;10(1):1-1.
25. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions.BMJ. 2016;355.
26. Sterne JA, Hernán MA, McAleenan A, Reeves BC, Higgins JP. Chapter 25: Assessing risk of bias in a non-randomized study. Cochrane. 2019.
27. Kumar S, Soldatos G, Ranasinha S, Teede H, Pallin M. Continuous glucose monitoring versus self-monitoring of blood glucose in the management of cystic fibrosis related diabetes: A systematic review and meta-analysis. J Cyst Fibros. 2022.
28. Islam N, Ayele HT, Yu OH, Douros A, Filion KB. Sulfonylureas and the risk of ventricular arrhythmias among people with type 2 diabetes: a systematic review of observational studies. Clin Pharmacol Ther. 2022.
29. Golden SH, Brown T, Yeh HC, Maruthur N, Ranasinghe P, Berger Z, et al. Table 6. Study quality of observational studies comparing insulin delivery or glucose monitoring methods for diabetes mellitus. In: Methods for insulin delivery and glucose monitoring: comparative effectiveness. Agency for Healthcare Research and Quality (US), Rockville (MD). 2012.
30. Hampson G, Towse A, Dreitlein WB, Henshall C, Pearson SD. Real-world evidence for coverage decisions: opportunities and challenges. J Comp Eff Res. 2018;7(12):1133-43.
31. Burns L, Le Roux N, Kalesnik-Orszulak R, Christian J, Hukkelhoven M, Rockhold F, et al. Real-World Evidence for Regulatory Decision-Making: Guidance From Around the World. Clin. Ther. 2022;44(3):420-37.
32. Beaulieu-Jones BK, Finlayson SG, Yuan W, Altman RB, Kohane IS, Prasad V, et al. Examining the use of real-world evidence in the regulatory process. Clin Pharmacol Ther. 2020;107(4):843-52.
33. Roberts MH, Ferguson GT. Real-World evidence: bridging gaps in evidence to guide payer decisions. PharmacoEconomics - Open. 2021;5(1):3-11.
34. Quigley JM, Thompson JC, Halfpenny NJ, Scott DA. Critical appraisal of nonrandomized studies—a review of recommended and commonly used tools. J Eval Clin Pract. 2019;25(1):44-52.
35. Buccheri RK, Sharifi C. Critical appraisal tools and reporting guidelines for evidence-based practice. Worldviews Evid Based Nurs. 2017;14(6):463-72.
36. Berger ML, Martin BC, Husereau D, Worley K, Allen JD, Yang W, et al. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value health. 2014;17(2):143-56.
37. Dziak JJ, Dierker LC, Abar B. The interpretation of statistical power after the data have been gathered. Curr Psychol. 2020;39(3):870-7.
38. Zhang Y, Hedo R, Rivera A, Rull R, Richardson S, Tu XM. Post hoc power analysis: is it an informative and meaningful analysis?. Gen Psychiatr. 2019;32(4).
39. Kim J, Seo BS. How to calculate sample size and why. Clin Orthop Surg. 2013 ;5(3):235-42.
40. Button KS, Ioannidis J, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365-76.
41. Bacchetti P. Small sample size is not the real problem. Nat Rev Neurosci. 2013;14(8):585.
42. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015;12(10):e1001885.

43. Whiting P, Wolff R, Mallett S, Simera I, Savović J. A proposed framework for developing quality assessment tools. Syst Rev. 2017;6(1):1-9.

44. Wang Z, Taylor K, Allman-Farinelli M, Armstrong B, Askie L, Ghersi D, et al. A systematic review: Tools for assessing methodological quality of human observational studies. Preprint at https://osf.io/preprints/metaarxiv/pnqmy (2019).

45. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015;4(1):1-22.

46. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Res Synth Methods. 2011;2(1):1-4.

# Appendices

**Appendix 1. Search strategy**

| |
|---|
| **Embase(**<br>('diabetes' OR 'diabetic' OR 'polyuria' OR 'diuresis' OR 'polydipsia' OR 'polygenic' OR 'metabolic syndrome*' OR 'dm' OR 'iddm' OR 'niddm' OR 'mody'):ti,ab,kw OR 'diabetes mellitus'/exp)<br>**AND(**<br>('glucose monitor*':ti,kw OR 'blood glucose':ti,kw OR 'blood sugar':ti,kw OR 'glucose management':ti,kw OR 'blood glucose monitoring'/exp OR 'glucose blood level'/de) AND ('system$' OR 'analyzer$' OR 'analyser$' OR 'sensor$' OR 'continuous' OR 'flash' OR 'real time' OR 'measurement' OR 'self' OR 'home' OR 'intermittent' OR 'pump*'):ti,kw) OR ('freestyle libre' OR 'fgm' OR 'flash gm' OR 'cgm' OR 'cgsm' OR 'csii' OR 'smbg' OR 'hbgm' OR 'bgsm'):ti,kw<br>**AND**<br>("a1c" OR "hemoglobin a1c" OR "h?ba1c" OR "hb?1c" OR "glycemic" OR "glycemia" OR "glycated" OR "hyperglycemia"):ti,ab,kw OR 'hemoglobin A1c'/exp OR 'glycemic control'/exp<br>**AND**<br>"article"/it<br>**NOT**<br>"animal"/exp NOT "human"/exp<br>**Filter**<br>Published between 2011 and 2021 |
| **PubMed**<br>("diabetes" [tiab] OR "diabetic" [tiab] OR "polyuria" [tiab] OR "diuresis" [tiab] OR "polydipsia" [tiab] OR "polygenic" [tiab] OR "metabolic syndrome*" [tiab] OR "dm" [tiab] OR "iddm" [tiab] OR "niddm" [tiab] OR "mody" [tiab] OR "Diabetes Mellitus"[Mesh])<br>**AND**<br>((("glucose monitor*" [ti] OR "blood glucose" [ti] OR "blood sugar" [ti] OR "glucose management" [ti] OR "blood glucose"[MeSH] OR "blood glucose self-monitoring"[MeSH]) AND ("system$" [ti] OR "analyzer$" [ti] OR "analyser$" [ti] OR "sensor$" [ti] OR "continuous" [ti] OR "flash" [ti] OR "real time" [ti] OR "measurement" [ti] OR "self"[ti] OR "home" [ti] OR "intermittent" [ti] OR "pump$" [ti] )) OR "FreeStyle Libre"[ti] OR "fgm" [ti] OR "flash gm"[ti] OR "cgm"[ti] OR "cgsm"[ti] OR "csii"[ti] OR "smbg"[ti] OR "hbgm" [ti] OR "bgsm"[ti])<br>**AND**<br>("a1c"[tiab] OR "hemoglobin a1c"[tiab] OR "h?ba1c"[tiab] OR "hb?1c"[tiab] OR "glycemic" [tiab] OR "glycemia"[tiab] OR "glycated"[tiab] OR "hyperglycemia"[tiab] OR "Glycated Hemoglobin A" [Mesh] OR "Glycemic Control" [Mesh])<br>**NOT**<br>("Animals"[Mesh] NOT "Humans"[Mesh])<br>**Filter**<br>Published in the past 10 years\ |

**Appendix 1. Continued**

| |
|---|
| **Scopus**<br>TITLE-ABS-KEY (“diabetes” OR “diabetic” OR “polyuria” OR “diuresis” OR “polydipsia” OR “polygenic” OR “metabolic syndrome*” OR “dm” OR “iddm” OR “niddm” OR “mody”)<br>**AND**<br>TITLE (((“glucose monitor*” OR “blood glucose” OR “blood sugar” OR “glucose management”) AND (“system$” OR “analyzer$” OR “analyser$” OR “sensor” OR “continuous” OR “flash” OR “real time” OR “measurement” OR “self” OR “home” OR “intermittent” OR “pump*” )) OR “FreeStyle Libre” OR "fgm" OR “flash gm” OR “cgm” OR “cgsm” OR “csii” OR “smbg” OR “hbgm” OR “bgsm” )<br>**AND**<br>TITLE-ABS-KEY (“a1c” OR “hemoglobin a1c” OR “h?ba1c” OR “hb?1c” OR “glycemic” OR “glycemia” OR “glycated” OR “hyperglycemia”)<br>**AND**<br>DOCTYPE (ar)<br>**AND NOT**<br>SUBJAREA(VETE)<br>**Filter**<br>Published between 2011 and 2021 |
| **Web of Science**<br>TS=(“diabetes” OR “diabetic” OR “polyuria” OR “diuresis” OR “polydipsia” OR “polygenic” OR “metabolic syndrome*” OR “dm” OR “iddm” OR “niddm” OR “mody”)<br>**AND**<br>TI=(((“glucose monitor*” OR “blood glucose” OR “blood sugar” OR “glucose management”) AND (“system*” OR “analyzer*” OR “analyser*” OR “sensor” OR “continuous” OR “flash” OR “real time” OR “measurement” OR “self” OR “home” OR “intermittent” OR “pump*” )) OR “FreeStyle Libre” OR "fgm" OR “flash gm” OR “cgm” OR “cgsm” OR “csii” OR “smbg” OR “hbgm” OR “bgsm” )<br>**AND**<br>TS=(“a1c” OR “hemoglobin a1c” OR “h?ba1c” OR “hb?1c” OR “glycemic” OR “glycemia” OR “glycated” OR “hyperglycemia”)<br>**NOT**<br>SU=Veterinary Sciences<br>**Filter**<br>Published in the past 10 years |

**Appendix 2. Reference list of 72 eligible studies**


1. Akturk HK, Giordano D, Champakanath A, Brackett S, Garg S, Snell-Bergeon J. Long-term real-life glycaemic outcomes with a hybrid closed-loop system compared with sensor-augmented pump therapy in patients with type 1 diabetes. Diabetes Obes Metab. 2020 Apr;22(4):583-9.
2. Anderson J, Attvall S, Sternemalm L, Pivodic A, Fahlén M, Hanås R, et al. Effect on glycemic control by short-and long-term use of continuous glucose monitoring in clinical practice. J Diabetes Sci Technol. 2011 Nov;5(6):1472-9.
3. Anjana RM, Kesavadev J, Neeta D, Tiwaskar M, Pradeepa R, Jebarani S, et al. A multicenter real-life study on the effect of flash glucose monitoring on glycemic control in patients with type 1 and type 2 diabetes. Diabetes Technol Ther. 2017 Sep 1;19(9):533-40.
4. Beato-Víbora PI, Quiros-Lopez C, Lázaro-Martín L, Martín-Frías M, Barrio-Castellanos R, Gil-Poch E, et al. Impact of sensor-augmented pump therapy with predictive low-glucose suspend function on glycemic control and patient satisfaction in adults and children with type 1 diabetes. Diabetes Technol Ther. 2018 Nov 1;20(11):738-43.
5. Cherubini V, Bonfanti R, Casertano A, De Nitto E, Iannilli A, Lombardo F, et al. Time in range in children with type 1 diabetes using treatment strategies based on nonautomated insulin delivery systems in the real world. Diabetes Technol Ther. 2020 Jul 1;22(7):509-15.
6. Deshmukh H, Wilmot EG, Gregory R, Barnes D, Narendran P, Saunders S, et al. Effect of flash glucose monitoring on glycemic control, hypoglycemia, diabetes-related distress, and resource utilization in the Association of British Clinical Diabetologists (ABCD) nationwide audit. Diabetes Care. 2020 Sep 1;43(9):2153-60.
7. Frontino G, Bonfanti R, Scaramuzza A, Rabbone I, Meschi F, Rigamonti A, et al. Sensor-augmented pump therapy in very young children with type 1 diabetes: an efficacy and feasibility observational study. Diabetes Technol Ther. 2012 Sep 1;14(9):762-4.
8. Gill M, Zhu C, Shah M, Chhabra H. Health care costs, hospital admissions, and glycemic control using a standalone, real-time, continuous glucose monitoring system in commercially insured patients with type 1 diabetes. J Diabetes Sci Technol. 2018 Jul;12(4):800-7.
9. Greve SV, Stilgren L. A pragmatic real-life study of flash glucose monitoring versus self-monitoring of blood glucose. Dan Med J. 2020 Jun 1;67(6):A07190404.
10. Gurnurkar S, Owens L, Chalise S, Vyas N. Evaluation of Hemoglobin A1c before and after initiation of continuous glucose monitoring in children with type 1 diabetes mellitus. J Pediatr Endocrinol Metab. 2021 Mar 1;34(3):311-7.
11. Hidefjäll P, Berg L. Patient controlled, off-label use of continuous glucose monitoring: real-world medical costs and effects of patient controlled sensor augmented pump therapy in adult patients type 1 diabetes. J Diabetes Sci Technol. 2021 May;15(3):575-81.
12. Katayama A, Tone A, Watanabe M, Teshigawara S, Miyamoto S, Eguchi J, et al. The hypoglycemia-prevention effect of sensor-augmented pump therapy with predictive low glucose management in Japanese patients with type 1 diabetes mellitus: a short-term study. Diabetol Int. 2020 Apr;11(2):97-104.
13. Kesavadev J, Vigersky R, Shin J, Pillai PB, Shankar A, Sanal G, et al. Assessing the therapeutic utility of professional continuous glucose monitoring in type 2 diabetes across various therapies: a retrospective evaluation. Adv Ther. 2017 Aug;34(8):1918-27.
14. Landau Z, Abiri S, Gruber N, Levy-Shraga Y, Brener A, Lebenthal Y, et al. Use of flash glucose-sensing technology (FreeStyle Libre) in youth with type 1 diabetes: AWeSoMe study group real-life observational experience. Acta Diabetol. 2018 Dec;55(12):1303-10.

15. Lepore G, Scaranna C, Corsi A, Dodesini AR, Trevisan R. Switching From Suspend-Before-Low Insulin Pump Technology to a Hybrid Closed-Loop System Improves Glucose Control and Reduces Glucose Variability: A Retrospective Observational Case–Control Study. Diabetes Technol Ther. 2020 Apr 1;22(4):321-5.

16. Lin J, Li X, Jiang S, Ma X, Yang Y, Zhou Z. Utilizing Technology-Enabled Intervention to Improve Blood Glucose Self-Management Outcome in Type 2 Diabetic Patients Initiated on Insulin Therapy: A Retrospective Real-World Study. Int J Endocrinol. 2020 Nov 10;2020.

17. Lou G, Larramona G, Montaner T, Barbed S. Effect of CGM in the HbA1c and Coefficient of Variation of glucose in a pediatric sample. Prim Care Diabetes. 2021 Apr 1;15(2):289-92.

18. Nana M, Moore SL, Ang E, Lee ZX, Bondugulapati LN. Flash glucose monitoring: Impact on markers of glycaemic control and patient-reported outcomes in individuals with type 1 diabetes mellitus in the real-world setting. Diabetes Res Clin Pract. 2019 Nov 1;157:107893.

19. Parkin CG, Graham C, Smolskis J. Continuous glucose monitoring use in type 1 diabetes: longitudinal analysis demonstrates meaningful improvements in HbA1c and reductions in health care utilization. J Diabetes Sci Technol. 2017 May;11(3):522-8.

20. Patrascioiu I, Quirós C, Ríos P, Ruíz M, Mayordomo R, Conget I, et al. Transitory beneficial effects of professional continuous glucose monitoring on the metabolic control of patients with type 1 diabetes. Diabetes Technol Ther. 2014 Apr 1;16(4):219-23.

21. Pepper GM, Steinsapir J, Reynolds K. Effect of short-term iPRO continuous glucose monitoring on hemoglobin A1c levels in clinical practice. Diabetes Technol Ther. 2012 Aug 1;14(8):654-7.

22. Ramirez-Rincon A, Hincapie-Garcia J, Arango CM, Aristizabal N, Castillo E, Hincapie G, et al. Clinical outcomes after 1 year of augmented insulin pump therapy in patients with diabetes in a specialized diabetes center in Medellín, Colombia. Diabetes Technol Ther. 2016 Nov 1;18(11):713-8.

23. Restrepo-Moreno M, Ramírez-Rincón A, Hincapié-García J, Palacio A, Monsalve-Arango C, Aristizabal-Henao N, et al. Maternal and perinatal outcomes in pregnant women with type 1 diabetes treated with continuous subcutaneous insulin infusion and real time continuous glucose monitoring in two specialized centers in Medellin, Colombia. J Matern Fetal Neonatal Med. 2018 Mar 19;31(6):696-700.

24. Sandig D, Grimsmann J, Reinauer C, Melmer A, Zimny S, Müller-Korbsch M, et al. Continuous glucose monitoring in adults with type 1 diabetes: real-world data from the German/Austrian prospective diabetes follow-up registry. Diabetes Technol Ther. 2020 Aug 1;22(8):602-12.

25. Sierra JA, Shah M, Gill MS, Flores Z, Chawla H, Kaufman FR, et al. Clinical and economic benefits of professional CGM among people with type 2 diabetes in the United States: analysis of claims and lab data. J Med Econ. 2018 Mar 4;21(3):225-30.

26. Stone MP, Agrawal P, Chen X, Liu M, Shin J, Cordero TL, et al. Retrospective analysis of 3-month real-world glucose data after the MiniMed 670G system commercial launch. Diabetes Technol Ther. 2018 Oct 1;20(10):689-92.

27. Swaney EE, McCombe J, Coggan B, Donath S, O'Connell MA, Cameron FJ. Has subsidized continuous glucose monitoring improved outcomes in pediatric diabetes?. Pediatr Diabetes. 2020 Nov;21(7):1292-300.

28. Tsur A, Cahn A, Israel M, Feldhamer I, Hammerman A, Pollack R. Impact of flash glucose monitoring on glucose control and hospitalization in type 1 diabetes: a nationwide cohort study. Diabetes Metab Res Rev. 2021 Jan;37(1):e3355.

29. Quispe BV, Frías MM, Martín MB, Valverde RY, Gómez MÁ, Castellanos RB. Effectiveness of MiniMed 640G with SmartGuard® System for prevention of hypoglycemia in pediatric patients with type 1 diabetes mellitus. Endocrinol Diabetes Nutr (Engl Ed). 2017 Apr 1;64(4):198-203.

30. Viñals C, Quirós C, Giménez M, Conget I. Real-life management and effectiveness of insulin pump with or without continuous glucose monitoring in adults with type 1 diabetes. Diabetes Ther. 2019 Jun;10(3):929-36.

31. Virdi N, Daskiran M, Nigam S, Kozma C, Raja P. The association of self-monitoring of blood glucose use with medication adherence and glycemic control in patients with type 2 diabetes initiating non-insulin treatment. Diabetes Technol Ther. 2012 Sep 1;14(9):790-8.

32. Battelino T, Liabat S, Veeze HJ, Castaneda J, Arrieta A, Cohen O. Routine use of continuous glucose monitoring in 10 501 people with diabetes mellitus.

33. Diabet Med. 2015 Dec;32(12):1568-74.

34. Gil-Ibáñez MT, Aispuru GR. Cost-effectiveness analysis of glycaemic control of a glucose monitoring system (FreeStyle Libre®) for patients with type 1 diabetes in primary health care of Burgos. Enferm Clin (Engl Ed). 2020 Mar 1;30(2):82-8.

35. Pastakia SD, Cheng SY, Kirui NK, Kamano JH. Dynamics, impact, and feasibility of self-monitoring of blood glucose in the rural, resource-constrained setting of western Kenya. Clin Diabetes. 2015 Jul 1;33(3):136-43.

36. Faulds ER, Zappe J, Dungan KM. Real-world implications of hybrid close loop (HCL) insulin delivery system. Endocr Pract. 2019 May 1;25(5):477-84.

37. Mulla BM, Noor N, James-Todd T, Isganaitis E, Takoudes TC, Curran A, et al. Continuous glucose monitoring, glycemic variability, and excessive fetal growth in pregnancies complicated by type 1 diabetes. Diabetes Technol Ther. 2018 Jun 1;20(6):413-9.

38. Addala A, Maahs DM, Scheinker D, Chertow S, Leverenz B, Prahalad P. Uninterrupted continuous glucose monitoring access is associated with a decrease in HbA1c in youth with type 1 diabetes and public insurance. Pediatr Diabetes. 2020 Nov 1;21(7):1301-9.

39. Argento NB, Nakamura K. Personal real-time continuous glucose monitoring in patients 65 years and older. Endocr Pract. 2014 Dec 1;20(12):1297-302.

40. Crăciun AE, Bala C, Crăciun C, Roman G, Georgescu C, Hâncu N. The Use of Continuous Glucose Monitoring System in Combination with Individualized Lifestyle and Therapeutic Recommandations on Glycemic Control of Type 2 Diabetes Patients. Rom J Diabetes Nutr Metab Dis. 2014 Dec 15;21(4):291-9.

41. DeSalvo DJ, Miller KM, Hermann JM, Maahs DM, Hofer SE, Clements MA, et al. Continuous glucose monitoring and glycemic control among youth with type 1 diabetes: International comparison from the T1D Exchange and DPV Initiative. Pediatr Diabetes. 2018 Nov;19(7):1271-5.

42. Joo EY, Lee JE, Kang HS, Park SG, Hong YH, Shin YL, et al. Frequency of self-monitoring of blood glucose during the school day is associated with the optimal glycemic control among Korean adolescents with type 1 diabetes. Diabetes Metab J. 2018 Dec;42(6):480-7.

43. Kesavadev J, Shankar A, Pillai PB, Krishnan G, Jothydev S. Cost-effective use of telemedicine and self-monitoring of blood glucose via Diabetes Tele Management System (DTMS) to achieve target glycosylated hemoglobin values without serious symptomatic hypoglycemia in 1,000 subjects with type 2 diabetes mellitus—A retrospective study. Diabetes Technol Ther. 2012 Sep 1;14(9):772-6.

44. Kim SK, Kim HJ, Kim T, Hur KY, Kim SW, Lee MK, et al. Effectiveness of 3-day continuous glucose monitoring for improving glucose control in type 2 diabetic patients in clinical practice. Diabetes Metab J. 2014 Dec 1;38(6):449-55.

45. Kristensen K, Ögge LE, Sengpiel V, Kjölhede K, Dotevall A, Elfvin A, et al. Continuous glucose monitoring in pregnant women with type 1 diabetes: an observational cohort study of 186 pregnancies. Diabetologia. 2019 Jul;62(7):1143-53.

46. Kröger J, Fasching P, Hanaire H. Three European retrospective real-world chart review studies to determine the effectiveness of flash glucose monitoring on HbA1c in adults with type 2 diabetes. Diabetes Ther. 2020 Jan;11(1):279-91.

47. Leinung M, Nardacci E, Patel N, Bettadahalli S, Paika K, Thompson S. Benefits of short-term professional continuous glucose monitoring in clinical practice. Diabetes Technol Ther. 2013 Sep 1;15(9):744-7.

48. Madeo B, Diazzi C, Granata AR, Ghoch ME, Greco C, Romano S, et al. Effect of a standard schema of self-monitoring blood glucose in patients with poorly controlled, non-insulin-treated type 2 diabetes mellitus: a controlled longitudinal study. J Popul Ther Clin Pharmacol. 2020 Sep 2; 27(2): 1-11.

49. Melmer A, Züger T, Lewis DM, Leibrand S, Stettler C, Laimer M. Glycaemic control in individuals with type 1 diabetes using an open source artificial pancreas system (OpenAPS). Diabetes Obes Metab. 2019 Oct;21(10):2333-7.

50. Moreno-Fernandez J, Pazos-Couselo M, González-Rodriguez M, Rozas P, Delgado M, Aguirre M, et al. Clinical value of flash glucose monitoring in patients with type 1 diabetes treated with continuous subcutaneous insulin infusion. Endocrinol Diabetes Nutr (Engl Ed). 2018 Dec 1;65(10):556-63.

51. Nefs G, Bazelmans E, Marsman D, Snellen N, Tack CJ, de Galan BE. RT-CGM in adults with type 1 diabetes improves both glycaemic and patient-reported outcomes, but independent of each other. Diabetes Res Clin Pract. 2019 Dec 1;158:107910.

52. Rose L, Klausmann G, Seibold A. Improving HbA1c control in type 1 or type 2 diabetes using flash glucose monitoring: a retrospective observational analysis in two German centres. Diabetes Ther. 2021 Jan;12(1):363-72.

53. Sherrill CH, Houpt CT, Dixon EM, Richter SJ. Effect of pharmacist-driven professional continuous glucose monitoring in adults with uncontrolled diabetes. J Manag Care Spec Pharm. 2020 May;26(5):600-9.

54. Sia HK, Kor CT, Tu ST, Liao PY, Wang JY. Self-monitoring of blood glucose in association with glycemic control in newly diagnosed non-insulin-treated diabetes patients: a retrospective cohort study. Sci Rep. 2021 Jan 13;11(1):1-9.

55. Van Dril E, Schumacher C. Impact of professional continuous glucose monitoring by clinical pharmacists in an ambulatory care setting. J Am Coll Clin Pharm. 2019 Dec;2(6):638-44.

56. Virdi NS, Lefebvre P, Parisé H, Duh MS, Pilon D, Laliberté F, et al. Association of self-monitoring of blood glucose use on glycated hemoglobin and weight in newly diagnosed, insulin-naïve adult patients with type 2 diabetes. J Diabetes Sci Technol. 2013 Sep;7(5):1229-42.

57. Wu Z, Luo S, Zheng X, Bi Y, Xu W, Yan J, et al. Use of a do-it-yourself artificial pancreas system is associated with better glucose management and higher quality of life among adults with type 1 diabetes. Ther Adv Endocrinol Metab. 2020 Aug;11:2042018820950146.

58. Aloi J, Bode BW, Ullal J, Chidester P, McFarland RS, Bedingfield AE, et al. Comparison of an electronic glycemic management system versus provider-managed subcutaneous basal bolus insulin therapy in the hospital setting. J Diabetes Sci Technol. 2017 Jan;11(1):12-6.

59. Li FF, Liu BL, Zhu HH, Li T, Zhang WL, Su XF, et al. Continuous glucose monitoring in newly diagnosed type 2 diabetes patients reveals a potential risk of hypoglycemia in older men. J Diabetes Res. 2017 Feb 8;2017.

60. Foster NC, Miller KM, Tamborlane WV, Bergenstal RM, Beck RW. Continuous glucose monitoring in patients with type 1 diabetes using insulin injections. Diabetes Care. 2016 Jun 1;39(6):e81-2.

61. Dadlani V, Kaur RJ, Stegall M, Xyda SE, Kumari K, Bonner K, et al. Continuous glucose monitoring to assess glycemic control in the first 6 weeks after pancreas transplantation. Clin Transplant. 2019 Oct;33(10):e13719.

62. Omer SH, Kumar M, Al-Gathradhi M, Vijayaraghavalu S. Evaluation of Self Monitoring of Blood Glucose on Therapeutic Outcome of Type 2 Diabetes in Female Patients From Aseer Diabetic Center, Abha, Kingdom of Saudi Arabia. Pharm. Glob. 2017 Jul 1;8(3):54-61.

63. Distiller LA, Cranston I, Mazze R. First clinical experience with retrospective flash glucose monitoring (FGM) analysis in South Africa: characterizing glycemic control with ambulatory glucose profile. J Diabetes Sci Technol. 2016 Nov;10(6):1294-302.

64. Gomez-Peralta F, Dunn T, Landuyt K, Xu Y, Merino-Torres JF. Flash glucose monitoring reduces glycemic variability and hypoglycemia: real-world data from Spain. BMJ Open Diabetes Res Care. 2020 Mar 1;8(1):e001052.

65. Hajime M, Okada Y, Mori H, Uemura F, Sonoda S, Tanaka K, et al. Hypoglycemia in blood glucose level in type 2 diabetic Japanese patients by continuous glucose monitoring. Diabetol Metab Syndr. 2019 Dec;11(1):1-9.

66. Polonsky WH, Fortmann AL. Impact of real-time continuous glucose monitoring data sharing on quality of life and health outcomes in adults with type 1 diabetes. Diabetes Technol Ther. 2021 Mar 1;23(3):195-202.

67. Crăciun CI, Crăciun AE, Rusu A, Bocşan CI, Hâncu N, Buzoianu AD. Increased glycemic variability in type 2 diabetes patients treated with insulin-a real-life clinical practice, continuous glucose monitoring (CGM) study. Rev Rom Med Lab. 2018 Jul 1;26(3):345-52.

68. Agiro A, Xie Y, Bowman K, DeVries A. Leveraging benefit design for better diabetes self-management and A1C control. Am J Manag Care. 2018 Feb 1;24:e30-6.

69. Tweden KS, Deiss D, Rastogi R, Addaguduru S, Kaufman FR. Longitudinal analysis of real-world performance of an implantable continuous glucose sensor over multiple sensor insertion and removal cycles. Diabetes Technol Ther. 2020 May 1;22(5):422-7.

70. Divan V, Greenfield M, Morley CP, Weinstock RS. Perceived burdens and benefits associated with continuous glucose monitor use in type 1 diabetes across the lifespan. J Diabetes Sci Technol. 2022 Jan;16(1):88-96.

71. Omer SH, Al Qahtani MA, Altieb AM, Awwad AA, Al-Gathradhi M, Vijayaraghavalu S. Positive impact of self-monitoring of blood glucose on diabetes management in male patients with type 2 diabetes from aseer diabetic center, Abha, Kingdom of Saudi Arabia. Pharm. Glob. 2015 Jul 1;6(3):1.

72. Choudhary P, Ramasamy S, Green L, Gallen G, Pender S, Brackenridge A, et al. Real-time continuous glucose monitoring significantly reduces severe hypoglycemia in hypoglycemia-unaware patients with type 1 diabetes. Diabetes Care. 2013 Dec 1;36(12):4160-2.

**Appendix 3. Details of study characteristics**

| Study | Diabetes | Region | Setting[a] | Intervention[b] | Comparator[b] | Outcome[c] | Design | Data source[d] | Sample size (Intervention) |
|---|---|---|---|---|---|---|---|---|---|
| Akturk2020 | Type-1 | US | Secondary | Closed-loop | SAPT | HbA1c & TIR | Cohort | EM(H)R | 127 |
| Anderson2011 | Type-1 | Sweden | Primary | CGM(NR) | Non-user | HbA1c & Hypoglycemia | Cohort | EM(H)R | 77 |
| Anjana2017 | Type-1&2 | India | Primary | CGM(FGM) | Non-user | HbA1c & FBG & Other | Case-control | EM(H)R | 2536 |
| Beato-Víbora2018 | Type-1 | Spain | Secondary | SAPT | NC | HbA1c & TIR & Hypoglycemia | Cohort | EM(H)R | 162 |
| Cherubini2020 | Type-1 | Italy | NR | CGM(rt-CGM) | CGM(is-CGM) | TIR | Cross-sectional | EM(H)R | 340 |
| Deshmukh2020 | All types | UK | Primary & Secondary | CGM(FGM) | Non-user | HbA1c & TIR & Hospitalization & Other | Crossover | Audit | 3182 |
| Frontino2012 | Type-1 | Italy | Secondary | SAPT | NC | HbA1c & Hypoglycemia & DKA & Other | Cohort | EM(H)R | 28 |
| Gill2018 | Type-1 | US | NR | CGM(rt-CGM) | Non-user | Hospitalization & Healthcare costs & Other | Cross-sectional | Admin | 1027 |
| Greve2020 | Type-1 | Denmark | Primary | CGM(FGM) | SMBG | HbA1c | Cohort | EM(H)R | 128 |
| Gurnurkar2020 | Type-1 | US | Secondary | CGM(NR) | Non-user | HbA1c | Crossover | EM(H)R | 90 |
| Hidefjäll2020 | Type-1 | Sweden | Secondary | SAPT | SMBG | HbA1c & Healthcare costs | Crossover | EM(H)R & PRO | 187 |
| Katayama2020 | Type-1 | Japan | Secondary | SAPT & CGM(rt-CGM) | SAPT & CGM(rt-CGM) | HbA1c & Hypoglycemia & SG & Other | Crossover | EM(H)R | 16 |
| Kesavadev2017 | Type-2 | India | Secondary | CGM (p-CGM) | Non-user | HbA1c & Hypoglycemia & Hyperglycemia & Other | Case-control | EM(H)R | 296 |

**Appendix 3. Continued**

| Study | Diabetes | Region | Setting[a] | Intervention[b] | Comparator[b] | Outcome[c] | Design | Data source[d] | Sample size (Intervention) |
|---|---|---|---|---|---|---|---|---|---|
| Landau2018 | Type-1 | Israel | Primary & Secondary | CGM(FGM) | NC | HbA1c & TIR & Hyperglycemia & DKA & Other | Cohort | EM(H)R | 71 |
| Lepore2020 | Type-1 | Italy | Secondary | Closed-loop | SAPT | HbA1c & Hyperglycemia & DKA & Other | Case-control | EM(H)R | 20 |
| Lin2020 | Type-2 | China | NR | SMBG & Telemedicine | NC | FBG & Other | Cohort | EM(H)R | 14085 |
| Lou2020 | Type-1 | Spain | Secondary | CGM(NR) | Non-user | HbA1c & Hypoglycemia & Glucose variability | Crossover | EM(H)R | 59 |
| Nana2019 | Type-1 | UK | Secondary | CGM(FGM) | Non-user | HbA1c & Hyperglycemia & Other | Crossover | EM(H)R & PRO | 90 |
| Parkin2017 | Type-1 | US | Primary & Secondary | CGM(rt-CGM) | SMBG | HbA1c & Hospitalization | Cohort | Admin | 187 |
| Patrascioiu2014 | Type-1 | Spain | Secondary | CGM (p-CGM) | NC | HbA1c & Hypoglycemia | Cohort | EM(H)R | 67 |
| Pepper2012 | Type-1&2 | US | Secondary | CGM(NR) | Non-user | HbA1c | Crossover | EM(H)R | 102 |
| Ramirez-Rincon2016 | Type-1&2 | Colombia | Secondary | CGM(rt-CGM) | NC | HbA1c & Hypoglycemia & Hospitalization & Hyperglycemia & Other | Cohort | EM(H)R | 183 |
| Restrepo-Moreno2018 | Type-1 | Colombia | NR | CGM(rt-CGM) | NC | HbA1c & Hospitalization & Other | Cross-sectional | EM(H)R | 14 |
| Sandig2020 | Type-1 | Germany & Austria | NR | CGM(is-CGM) | CGM(rt-CGM) | HbA1c & TIR & Other | Cross-sectional | Registry | 185 |
| Sierra2018 | Type-2 | US | NR | CGM (p-CGM) | Non-user | HbA1c & Healthcare costs | Crossover | Admin | 5677 |

**Appendix 3. Continued**

| Study | Diabetes | Region | Setting[a] | Intervention[b] | Comparator[b] | Outcome[c] | Design | Data source[d] | Sample size (Intervention) |
|---|---|---|---|---|---|---|---|---|---|
| Stone2018 | Type-1 | US | NR | CGM(NR) | NC | TIR & SG & Other | Cohort | EM(H)R | 3141 |
| Swaney2020 | Type-1 | Australia | Secondary | CGM(NR) | Non-user | HbA1c | Crossover | EM(H)R | 341 |
| Tsur2020 | Type-1 | Israel | Primary & Secondary | CGM(FGM) | NC | HbA1c & Hypoglycemia & Hospitalization & Other | Cohort | EM(H)R | 2682 |
| Quispe2017 | Type-1 | Spain | Secondary | CGM(NR) | NC | HbA1c & Hypoglycemia & Hyperglycemia & FBG & Other | Crossover | EM(H)R | 21 |
| Vinals2019 | Type-1 | Spain | Secondary | SAPT | Non-user | HbA1c & Other | Case-control | EM(H)R | 40 |
| Virdi2012 | Type-2 | US | Secondary | SMBG | Non-user | HbA1c & Other | Crossover | Admin | 2744 |
| Battelino2016 | Type-1&2 | Western Europe & Israel & Canada | Primary & Secondary | SAPT | Non-user | HbA1c & Hypoglycemia & Glucose variability | Cohort | EM(H)R | 7916 |
| Gil-Ibáñez2020 | Type-1 | Spain | NR | CGM(FGM) | Non-user | Hypoglycemia | Crossover | Registry | 23 |
| Pastakia2015 | All types | Kenya | Primary | SMBG | NC | HbA1c | Cohort | EM(H)R | 137 |
| Faulds2018 | Type-1 | US | Secondary | Closed-loop | NC | HbA1c & TIR & Other | Cohort | EM(H)R | 34 |
| Mulla2018 | Type-1 | US | NR | CGM(NR) | NC | HbA1c & Other | Cohort | EM(H)R | 41 |
| Addala2020 | Type-1 | US | Primary | CGM(NR) | Non-user | HbA1c & TIR & Hypoglycemia & Hyperglycemia | Cohort | EM(H)R | 115 |
| Argento2014 | Type-1&2 | US | Primary | CGM(rt-CGM) | Non-user | HbA1c & Hyperglycemia | Crossover | EM(H)R | 39 |

**Appendix 3. Continued**

| Study | Diabetes | Region | Setting[a] | Intervention[b] | Comparator[b] | Outcome[c] | Design | Data source[d] | Sample size (Intervention) |
|---|---|---|---|---|---|---|---|---|---|
| Crăciun2014 | Type-2 | Romania | Primary | CGM(NR) | NC | HbA1c & Glucose variability | Cohort | EM(H)R | 28 |
| DeSalvo2018 | Type-1 | US & German & Austrian | NR | CGM(rt-CGM) or CGM(is-CGM) | NC | HbA1c | panel | Registry | 29007 (year 2011) & 29150 (year 2016) |
| Joo2018 | Type-1 | Korea | NR | SMBG | NC | HbA1c & Other | Cross-sectional | EM(H)R | 61 |
| Kesavadev2012 | Type-2 | India | Secondary | SMBG | NC | HbA1c & Healthcare costs & FBG & Other | Cohort | EM(H)R | 1000 |
| Kim2014 | Type-2 | Korea | Primary | CGM(NR) | Non-user | HbA1c | Case-control | EM(H)R | 65 |
| Kristensen2019 | Type-1 | Sweden | NR | CGM(rt-CGM) | CGM(is-CGM) | HbA1c & TIR & Other | Cohort | EM(H)R | 92 |
| Kröger2020 | Type-2 | Austria & France & Germany | Secondary | CGM(FGM) | NC | HbA1c | Cohort | EM(H)R | 363 |
| Leinung2013 | Type-1&2 | US | Primary | CGM (p-CGM) | NC | HbA1c & Hypoglycemia | Cohort | EM(H)R | 121 |
| Madeo2020 | Type-2 | Italy | Secondary | SMBG | Non-user | HbA1c & FBG | Case-control | EM(H)R | 27 |
| Melmer2019 | Type-1 | US | NR | Closed-loop | SAPT | HbA1c & TIR & Other | Crossover | EM(H)R | 34 |
| Moreno-Fernandez2018 | Type-1 | Spain | Secondary | CGM(FGM) | SMBG | HbA1c | Cohort | EM(H)R | 18 |
| Nefs2019 | Type-1 | Netherlands | Secondary | CGM(rt-CGM) | NC | HbA1c & Hypoglycemia | Cohort | EM(H)R | 54 |
| Rose2021 | Type-1&2 | Germany | Secondary | CGM(FGM) | NC | HbA1c | Cohort | EM(H)R | 307 |
| Sherrill2020 | All types | US | Primary | CGM (p-CGM) | NC | HbA1c & Other | Cohort | EM(H)R | 315 |
| Sia2021 | Type-2 | China | Secondary | SMBG | Non-user | HbA1c | Cohort | EM(H)R | 1047 |

**Appendix 3. Continued**

| Study | Diabetes | Region | Setting[a] | Intervention[b] | Comparator[b] | Outcome[c] | Design | Data source[d] | Sample size (Intervention) |
|---|---|---|---|---|---|---|---|---|---|
| Van Dril2019 | Type-1&2 | US | Primary | CGM (p-CGM) | Non-user | HbA1c & Other | Crossover | EM(H)R | 29 |
| Virdi2013 | Type-2 | US | NR | SMBG | Non-user | HbA1c & Other | Cohort | EM(H)R & Admin | 589 |
| Wu2020 | Type-1 | China | NR | CGM(rt-CGM) | Non-user | HbA1c & Glucose variability & Other | Crossover | Registry & PRO | 15 |
| Aloi2017 | All types | US | Secondary | eGMS | Non-user | HbA1c & Hypoglycemia & Other | Crossover | EM(H)R | 993 |
| Li2017 | Type-2 | China | Secondary | CGM(retro-CGM) | NC | Hyperglycemia & Other | Cohort | EM(H)R | 106 |
| Foster2016 | Type-1 | US | NR | CGM(NR) | NC | HbA1c | Cross-sectional | Registry | 17731 |
| Dadlani2019 | Type-1 | US | Secondary | CGM(NR) | NC | Glucose variability | Cohort | EM(H)R | 26 |
| Omer2017 | Type-2 | Saudi Arabia | Secondary | SMBG | Non-user | HbA1c & Other | Case-control | EM(H)R | 157 |
| Distiller2016 | Type-1&2 | South Africa | Primary | CGM(FGM) | NC | HbA1c | Cross-sectional | EM(H)R | 50 |
| Gomez-Peralta2019 | Type-1 | Spain | NR | CGM(FGM) | NC | HbA1c & Hypoglycemia & Hyperglycemia & Other | Cross-sectional | EM(H)R | 22949 |
| Hajime2019 | Type-2 | Japan | Secondary | CGM(NR) | NC | HbA1c & Other | Cohort | EM(H)R | 293 |
| Polonsky2021 | Type-1 | US | NR | CGM(rt-CGM) | NC | HbA1c & Hypoglycemia & Other | Cross-sectional | EM(H)R | 302 |
| Crăciun2018 | Type-2 | Romania | NR | CGM(NR) | NC | Glucose variability & Other | Cross-sectional | EM(H)R | 95 |

**Appendix 3. Continued**

| Study | Diabetes | Region | Setting[a] | Intervention[b] | Comparator[b] | Outcome[c] | Design | Data source[d] | Sample size (Intervention) |
|---|---|---|---|---|---|---|---|---|---|
| Agiro2018 | All types | US | NR | SMBG | NC | HbA1c & Healthcare costs & Other | Cohort | Admin | 7155 |
| Tweden2020 | All types | Europe & South Africa | NR | CGM(NR) | NC | Hypoglycemia & Hyperglycemia & SG & Other | Cohort | EM(H)R | 945 |
| Divan2020 | Type-1 | US | Secondary | CGM(NR) | Non-user | HbA1c & Other | Cross-sectional | Registry & PRO | 702 |
| Omer2015 | Type-2 | Saudi arabia | Secondary | SMBG | SMBG | HbA1c | Case-control | EM(H)R | 200 |
| Choudhary2013 | Type-1 | UK | NR | CGM(NR) | NC | HbA1c & Other | Cohort | Audit | 35 |
| Qayyum2016 | Type-1&2 | Singapore | Secondary | CGM(NR) | NC | HbA1c | Cohort | EM(H)R | 60 |

[a] NR indicates the information was not reported;

[b] CGM(FGM) indicates flash glucose monitoring; CGM(rt-CGM), real-time continuous glucose monitoring; CGM (p-CGM), professional continuous glucose monitoring; CGM(is-CGM), intermittently scanned glucose monitoring; CGM(retro-CGM),retrospective continuous glucose monitoring; CGM(NR), continuous glucose monitoring without reporting the subtypes; SAPT, sensor augmented pump therapy ; SMBG, self-monitoring of blood glucose; eGMS, electronic glycemic management system; NC, no comparator; Non-user, non-user group.

[c] HbA1c indicates hemoglobine A1c; TIR, time in range; FBG, fasting blood glucose; DKA, diabetic ketoacidosis; SG, sensor glucose values; Other indicates outcomes not mentioned by this appendix;

[d] EM(H)R indicates electronic medical or health records; Admin, healthcare administrative data; PRO, patient reported outcomes.

**Appendix 4. Number of studies that fulfilled signaling questions (the ROBINS-I), stratified by publication years**

| Criterion (the ROBINS-I checklist) | Applicable | 2011-2017 | | | 2018-2019 | | | 2020-2021 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Y/PY | N/PN | NI | Y/PY | N/PN | NI | Y/PY | N/PN | NI |
| ***Bias due to confounding*** | | | | | | | | | | |
| 1.1 Is there potential for confounding of the effect of intervention in this study? | 72 (100%) | 26 (100%) | 0 | 0 | 22 (100%) | 0 | 0 | 24 (100%) | 0 | 0 |
| 1.2 Was the analysis based on splitting participants' follow up time according to intervention received? | 72 (100%) | 8 (31%) | 18 (69%) | 0 | 3 (14%) | 19 (86%) | 0 | 10 (42%) | 14 (58%) | 0 |
| 1.3 Were intervention discontinuations or switches likely to be related to factors that are prognostic for the outcome? | 21 (29%) | 5 (62%) | 1 (13%) | 2 (25%) | 1 (34%) | 0 | 2 (66%) | 2 (20%) | 3 (30%) | 5 (50%) |
| 1.4 Did the authors use an appropriate analysis method that controlled for all the important confounding domains? | 55 (76%) | 8 (42%) | 9 (47%) | 2 (11%) | 11 (58%) | 7 (37%) | 1 (5%) | 13 (76%) | 4 (24%) | 0 |
| 1.5 Were confounding domains that were controlled for measured validly and reliably by the variables available in this study? | 35 (49%) | 4 (40%) | 1 (10%) | 5 (50%) | 7 (59%) | 3 (25%) | 2 (16%) | 10 (77%) | 3 (23%) | 0 |
| 1.6 Did the authors control for any post-intervention variables that could have been affected by the intervention? | 55 (76%) | 1 (6%) | 13 (68%) | 5 (26%) | 0 | 15 (79%) | 4 (21%) | 3 (18%) | 7 (41%) | 7 (41%) |
| 1.7 Did the authors use an appropriate analysis method that adjusted for all the important confounding domains and for time varying confounding? | 17 (24%) | 4 (57%) | 3 (43%) | 0 | 2 (67%) | 1 (33%) | 0 | 4 (57%) | 3 (43%) | 0 |
| 1.8 Were confounding domains that were adjusted for measured validly and reliably by the variables available in this study? | 10 (14%) | 4 (100%) | 0 | 0 | 1 (50%) | 1 (50%) | 0 | 3 (75%) | 1 (25%) | 0 |

**Appendix 4. Continued**

| Criterion (the ROBINS-I checklist) | Applicable | 2011-2017 | | | 2018-2019 | | | 2020-2021 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Y/PY | N/PN | NI | Y/PY | N/PN | NI | Y/PY | N/PN | NI |
| ***Bias in selection of participants into the study*** | | | | | | | | | | |
| 2.1 Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention? | 72 (100%) | 0 | 25 (96%) | 1 (4%) | 1 (5%) | 21 (95%) | 0 | 6 (25%) | 18 (75%) | 0 |
| 2.2 Were the postintervention variables that influenced selection likely to be associated with intervention? | 8 (11%) | 1 (100%) | 0 | 0 | 0 | 1 (100%) | 0 | 5 (83%) | 1 (17%) | 0 |
| 2.3 Were the postintervention variables that influenced selection likely to be influenced by the outcome or a cause of the outcome? | 6 (8%) | 0 | 1 (100%) | 0 | 0 | 0 | 0 | 5 (100%) | 0 | 0 |
| 2.4 Do start of follow-up and start of intervention coincide for most participants? | 72 (100%) | 16 (61%) | 7 (27%) | 3 (12%) | 16 (73%) | 2 (9%) | 4 (18%) | 16 (67%) | 5 (21%) | 3 (12%) |
| 2.5 Were adjustment techniques used that are likely to correct for the presence of selection biases? | 28 (39%) | 7 (64%) | 4 (36%) | 0 | 2 (33%) | 4 (67%) | 0 | 9 (82%) | 2 (18%) | 0 |
| ***Bias in classification of interventions*** | | | | | | | | | | |
| 3.1 Were intervention groups clearly defined? | 72 (100%) | 22 (85%) | 4 (15%) | 0 | 21 (95%) | 1 (5%) | 0 | 23 (96%) | 1 (4%) | 0 |
| 3.2 Was the information used to define intervention groups recorded at the start of the intervention? | 72 (100%) | 26 (100%) | 0 | 0 | 22 (100%) | 0 | 0 | 23 (96%) | 0 | 1 (4%) |
| 3.3 Could classification of intervention status have been affected (blinded) by knowledge of the outcome or risk of the outcome? | 72 (100%) | 3 (11%) | 21 (81%) | 2 (8%) | 0 | 22 (100%) | 0 | 1 (4%) | 21 (88%) | 2 (8%) |

**Appendix 4. Continued**

| Criterion (the ROBINS-I checklist) | Applicable | 2011-2017 | | | 2018-2019 | | | 2020-2021 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Y/PY | N/PN | NI | Y/PY | N/PN | NI | Y/PY | N/PN | NI |
| ***Bias due to deviations from intended interventions*** | | | | | | | | | | |
| 4.1 Were there deviations from the intended intervention beyond what would be expected in usual practice | 72 (100%) | 0 | 26 (100%) | 0 | 0 | 22 (100%) | 0 | 0 | 24 (100%) | 0 |
| 4.2 Were these deviations from intended intervention unbalanced between groups and likely to have affected the outcome? | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***Bias due to missing data*** | | | | | | | | | | |
| 5.1 Were outcome data available for all, or nearly all, participants? | 72 (100%) | 21 (81%) | 5 (19%) | 0 | 21 (95%) | 1 (5%) | 0 | 22 (92%) | 2 (8%) | 0 |
| 5.2 Were participants excluded due to missing data on intervention status? | 72 (100%) | 10 (38%) | 13 (50%) | 3 (12%) | 7 (32%) | 14 (64%) | 1 (4%) | 10 (42%) | 13 (54%) | 1 (4%) |
| 5.3 Were participants excluded due to missing data on other variables needed for the analysis? | 72 (100%) | 8 (31%) | 14 (54%) | 4 (15%) | 8 (36%) | 14 (64%) | 0 | 10 (42%) | 10 (42%) | 4 (16%) |
| 5.4 Are the proportion of participants and reasons for missing data similar across interventions? | 34 (47%) | 5 (36%) | 0 | 9 (64%) | 6 (75%) | 0 | 2 (25%) | 6 (50%) | 0 | 6 (50%) |
| 5.5 Is there evidence that results were robust to the presence of missing data? | 34 (47%) | 3 (21%) | 6 (43%) | 5 (36%) | 4 (50%) | 4 (50%) | 0 | 4 (34%) | 6 (50%) | 2 (16%) |
| ***Bias in measurement of outcomes*** | | | | | | | | | | |
| 6.1 Could the outcome measure have been influenced by knowledge of the intervention received? | 72 (100%) | 6 (23%) | 19 (73%) | 1 (4%) | 3 (14%) | 18 (82%) | 1 (4%) | 4 (17%) | 20 (83%) | 0 |

**Appendix 4. Continued**

| Criterion (the ROBINS-I checklist) | Applicable | 2011-2017 | | | 2018-2019 | | | 2020-2021 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Y/PY | N/PN | NI | Y/PY | N/PN | NI | Y/PY | N/PN | NI |
| ***Bias in measurement of outcomes*** | | | | | | | | | | |
| 6.2 Were outcome assessors aware of the intervention received by study participants? | 72 (100%) | 26 (100%) | 0 | 0 | 21 (95%) | 1 (5%) | 0 | 24 (100%) | 0 | 0 |
| 6.3 Were the methods of outcome assessment comparable across intervention groups? | 72 (100%) | 24 (92%) | 1 (4%) | 1 (4%) | 20 (91%) | 1 (5%) | 1 (4%) | 23 (96%) | 0 | 1 (4%) |
| 6.4 Were any systematic errors in measurement of the outcome related to intervention received? | 72 (100%) | 6 (23%) | 15 (58%) | 5 (19%) | 4 (19%) | 15 (68%) | 3 (13%) | 4 (17%) | 16 (67%) | 4 (16%) |
| ***Bias in selection of the reported result*** | | | | | | | | | | |
| 7.1 Is the reported effect estimate likely to be selected, on the basis of the results, from multiple outcome measurements within the outcome domain? | 72 (100%) | 5 (20%) | 17 (65%) | 4 (15%) | 3 (14%) | 19 (86%) | 0 | 6 (25%) | 17 (71%) | 1 (4%) |
| 7.2 Is the reported effect estimate likely to be selected, on the basis of the results, from multiple analyses of the intervention-outcome relationship? | 72 (100%) | 2 (8%) | 23 (88%) | 1 (4%) | 0 | 21 (95%) | 1 (4%) | 2 (9%) | 20 (83%) | 2 (8%) |
| 7.3 Is the reported effect estimate likely to be selected, on the basis of the results, from different subgroups? | 72 (100%) | 6 (23%) | 16 (62%) | 4 (15%) | 6 (27%) | 14 (64%) | 2 (9%) | 5 (21%) | 18 (75%) | 1 (4%) |

Data are presented as n(%) of studies that fulfilled each signaling question. [a]According to algorithms defined by the ROBINS-I, some signaling questions are not necessary to judge, depending on the results of the previous criteria. "Applicable" indicates n(%) of studies that need to be judged for each signaling question; [b]"Y/PY" indicates Yes or Probably Yes; "N/PN" indicates No or Probably No; "NI" indicates no information.

4

Chapter 5

# Tools for Assessing Quality of Studies Investigating Health Interventions using Real-world Data: a Literature Review and Content Analysis

Li Jiu, Michiel Hartog, Junfeng Wang, Rick A Vreman, Olaf H Klungel, Aukje K Mantel-Teeuwisse, Wim G Goettsch

*BMJ open. 2024;14(2):e075173*

# Abstract

## Background

Non-randomized studies of interventions (NRSIs), a type of real-world studies, have become increasingly useful for decision-making in clinical and health technology assessment (HTA) settings. Given quality concerns, NRSIs need to be rigorously appraised, and this rationalizes the development and use of tools for assessing quality of such studies. However, the increased numbers of appraisal tools and great heterogeneity in how quality items are addressed among the tools have posed challenges on tool selection. Hence, we aimed to identify existing appraisal tools for NRSIs, and to compare criteria the tools provide at the quality-item level.

## Methods

We conducted a targeted search of appraisal tools for NRSIs published from 2002 through three approaches: search of journal articles in Medline, snowballing search of reviews on appraisal tools, and grey literature search on websites of HTA agencies. Then, we conducted a content analysis to summarize quality items from identified tools. Tools for methodological quality and reporting were analyzed separately, using NVIVO12.

## Results

From the 230 tools identified in this review, 49 tools met inclusion criteria and were included for the content analysis. concerns regarding the quality of NRSI were categorized into eight domains and 26 items. The RTI Item Bank and STROBE were the most comprehensive tools for methodological quality and reporting respectively, as they addressed (n=20;17) and sufficiently described (n=18;13) the highest number of items. However, none of the tools covered all items. The items least addressed for methodological quality included outcome selection, outcome definition, and ethical approval, and for reporting included intervention selection, intervention measurement, and length of follow-up.

## Conclusions

Most of the appraisal tools have their own strengths, but none of them could address all quality concerns relevant to NRSIs. Even the most comprehensive tools can be complemented by several tools. We suggest decision-makers, researchers, and tool developers consider the quality-item level heterogeneity, when selecting a tool or identifying a research gap.

## Introduction

Real world data (RWD) generally refer to data collected during routine clinical practice, but their definitions could vary in settings (1). According to Makady et al., one of the RWD definitions is data collected without interference with treatment assignment (1). RWD that fit this definition are normally analyzed in non-randomized studies of interventions (NRSIs), which estimate effectiveness of a health intervention without randomizing intervention groups (2,3).

NRSIs provide evidence on clinical and cost-effectiveness of health interventions for decision-making, in clinical and health technology assessment (HTA) settings (4-9). For example , NRSIs could inform clinicians on what diagnosis or treatment strategies to adopt (4,5). Also, with NRSIs, HTA agencies could gain more certainty on validity of evidence from randomized controlled trials (RCTs), when deciding on which health intervention to reimburse and on which pricing strategy to adopt (6,7). Also, HTA stakeholders could exploit NRSIs to evaluate highly innovative or complex interventions, for which RCTs may be considered infeasible or unethical (8,9). Generally speaking, NRSIs have become increasingly useful, as they complement and sometimes replace RCTs, when RCTs are scarce or even infeasible to conduct (2,10).



However, the usefulness of NRSIs is often questioned due to quality concerns, in terms of risk of bias (RoB) and reporting. According to the Cochrane Handbook, NRSIs have higher RoB than RCTs, and are vulnerable to various types of bias, such as confounding, selection, and information bias (11). Also, the Professional Society for Health Economics and Outcomes Research (ISPOR) published a report in 2020, which stated that insufficient reporting on how a NRSI was generated was a major barrier for decision-makers to adopt NRSIs (12).

To address NRSI's quality concerns and to build decision-makers' confidence, NRSIs need to be rigorously appraised, and this rationalizes the development and use of appraisal tools. According to systematic reviews of appraisal tools for NRSIs, tens of tools have been developed in the past five decades (13-15). The growing number of tools has then brought a new challenge to users: how to select the best tool. To address this challenge, previous reviews have summarized quality items (i.e. a group of criteria or signaling questions for methodological quality or reporting), and compared whether existing tools addressed these items (13-15). Some example items include "measurement of outcomes", "loss to follow-up bias", "inclusion and exclusion criteria of target population", "sampling strategies to correct selection bias", etc (13). In addition, these reviews provided some general recommendations on tool selection, such as referring

to multiple tools for quality appraisal (14). However, information is still lacking on to what extent the tools address each quality item, and the heterogeneity of tools at the quality-item level. To take outcome measurement as an example, the Academy of Nutrition and Dietetics Quality Criteria (ANDQ) checklist mentions that outcomes should be measured with "standard, valid, and reliable data collection instruments, tests, and procedures" and "at an appropriate level of precision" (16). In contrast, the Good ReseArch for Comparative Effectiveness (GRACE) checklist considers the "valid and reliable" measurement as "objective rather than subject to clinical judgment" (17); while the Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I) checklist interprets the "standard" way as "comparable across study groups", and "valid and reliable" as low detection bias without "systematic errors" in outcome measurement (18). In summary, the heterogeneity in level of detail with which a tool addresses a quality item and the heterogeneity in content and format of signaling questions can pose a challenge when tools are selected, or even merged.

Hence, our study aimed to summarize and compare signaling questions or criteria in the tools provided at the quality-item level, through a content analysis. This research was performed as part of the HTx project (19). The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825162.

## Methods

### Protocol

To ensure credibility of the review and the content analysis, we registered a study protocol in the OSF registry (registration DOI: https://doi.org/10.17605/OSF.IO/KCSGX) on June 30 2022. The OSF registry is an online repository that accepts registration of all types of research projects, including reviews and content analyses (20).

### Scope

In our study, appraisal tools refer to tools, guidelines, instruments, or standards that provide guidance on how to report or assess any quality concern of NRSIs. NRSIs, according to the Cochrane Handbook, refer to any quantitative study estimating the effectiveness of an intervention without randomization to allocate patients to intervention groups (1). According to Makady et al., data collected in such NRSIs belongs to the second category of real-world data, i.e., those collected without interference with treatment assignment, patient monitoring or follow-up, or selection of study population (1).

## Search strategy

To identify appraisal tools for NRSIs from various potential sources, we adopted three approaches. A diagram illustrating how the three approaches complemented each other is shown in Appendix 1.

### *Database search*

In the first approach, we conducted a systematic review to identify articles on appraisal tools, through a database search using Medline. Since D'Andrea et al. (2021) have already conducted a systematic review to identify appraisal tools for all types of non-randomized studies published before November 2019 (13), we updated their review by searching for articles published between November 2019 and April 2022, with their strings.



### *Snowballing*

In the second approach, we searched for published reviews on appraisal tools for NRSIs. To identify all published reviews, we adopted a snowballing approach described by Wohlin (2016) (21). Snowballing refers to using the citations of articles to identify additional articles, and it is considered a good extension of a database search (21). To implement the snowballing approach, three researchers (LJ,MH, and JW) first conducted a pilot search of articles using Google Scholar, reviewed full-text, judged eligibility through a group discussion, then identified three reviews (i.e., those by D'Andrea et al. (2021) (13), Quigley et al. (2019) (14), and Faria et al. (2015) (15)). Next, the three reviews were used as a starting set, and were uploaded to the website Connected Papers, which provides an online tool for snowballing (22). With each uploaded review, Connected Papers analyzed approximately 50,000 articles, and finally returned 40 articles with the highest level of similarity, based on factors such as overlapping citations. After judging eligibility of the returned articles, eligible articles were uploaded to the website Connected Papers for a second round of snowballing.

### *Grey literature*

In the third approach, we searched for grey literature on the websites of European HTA agencies. Our rationale was that some appraisal tools may exist in the format of grey literature, such as agency reports and technical support documents. The list of European HTA agencies was derived from the International Network of Agencies for Health Technology Assessment (INAHTA) (23). On each agency website, two researchers (MH and LJ) independently searched for grey literature with four concepts respectively: "quality", "risk of bias"," appraisal", and "methodology". For each concept, only the first 10 hits sorted by relevance, if optional, were included (i.e. a maximum of 40 hits for each website).

### Eligibility criteria for articles and grey literature to identify relevant tools

An article or grey literature document was included if it described one or more appraisal tools. It was excluded if it only described tools for RCTs or only described tools for diagnostic, prognostic, qualitative, or secondary studies (e.g. systematic reviews and cost-effectiveness analyses). We only included articles identified through the database search and snowballing if published in English, while included grey literature could be published in all languages, as many HTA agencies tend to only use languages of their nations. Relevant documents obtained through this approach were translated using Google Translate.

### The process of identifying studies and appraisal tools

Two researchers (MH and LJ) independently scanned all titles and abstract of the identified hits, then reviewed the full-text with Rayyan (24) and Excel. After identifying the eligible studies, one researcher (MH) extracted the name of the tools, and downloaded them by tracking study citations. A pilot search with Google was conducted to ensure we downloaded the most up-to-date version. Next, two researchers (MH and LJ) independently reviewed full-text and judged eligibility of the tools. An appraisal tool was included if it (1) was designed for non-randomized studies, (2) was used for assessing either methodological quality or reporting, and (3) was developed or updated after 2002. A tool was excluded if it was designed for non-randomized studies of exposures which were not controlled by investigators (e.g. diets). All discrepancies were solved through discussion among the three researchers (MH, LJ, and JW).

### Data collection & Content analysis

One researcher (MH) extracted tool characteristics using a pre-specified Excel form. The data items included publication year, tool format (e.g., checklist or rating scale), targeted study design (e.g. all NRSIs, cohort studies, etc.), target interventions (e.g. all or surgical interventions), originality (i.e. whether a tool was developed based on an existing tool), and scope. The scope referred to whether the tools were designed for assessing methodological quality (e.g. risk of bias and external validity) and/or for ensuring adequate reporting of research details that could be used for assessing methodological quality (25).

For the content analysis, we adopted both deductive and inductive coding techniques (26). First, we derived a list of candidate quality items from the three reviews, the starting set for the snowballing (13-15). Then, in a pilot coding process, we reviewed all identified appraisal tools, and judged whether a candidate quality item was described. After the pilot coding, we summarized signaling questions or criteria that were not covered by the candidate items, and coded them as new items. After updating the list of candidate

items, three researchers (JW, LJ, and MH) finalized the items in four group meetings. During the meetings, we merged items with overlapping content, split items containing too much content, and renamed items so they could be self-explanatory.

To score whether and to what extent a quality item was described by a tool, we again reviewed all identified tools. If an item was described by a tool in one or several signaling questions, we judged whether the question(s) was related to methodological quality, reporting, or both, independently of what original studies claimed to be. Additionally, we judged whether an item was described sufficiently or briefly. A description was scored as "brief", if the corresponding signaling question(s) did not explain how to improve or assess methodological quality or specify elements needed for reporting. For example, "outcomes should be measured appropriately" or "outcome measurement should be adequately described" are "brief" descriptions, if no additional explanations were provided. The scoring process was independently conducted by two researchers (LJ and MH) using NVivo12, and all discrepancies were solved through discussion between the two.



## Results

### Tool selection

As shown in Figure 1, we identified 1738 articles after removing duplicates, and excluded 1645 articles after subsequently reviewing titles, abstracts, and full-text. From the remaining 27 eligible studies, we identified 417 appraisal tools. After removing duplicates and reviewing full-texts, we included 49 tools which met our criteria. References of the included studies and appraisal tools are shown in Appendix 2 and 3, respectively.

### Characteristics of appraisal tools

As shown in Table 1, 18 (37%) tools were published between 2002 and 2010, while 31 (63%) tools were published thereafter. Among these, 30 (61%), 6 (12%), and 5 (10%) tools were designed for addressing methodological quality, reporting, and both, respectively, while 7 (14%) tools did not report intended use of the tools. About three quarters of the tools were designed for all types of NRSIs, while others were designed for one or several NRSI types, such as cohort (16%) and case-control studies (16%). Regarding sources, 44 (90%) tools were described in articles that developed a tool, in grey literature (e.g. online checklist or report), or in both, while the other five tools were extended from existing tools, when researchers conducted systematic reviews on non-randomized studies. Finally, 9 (18%) tools were designed for specific interventions or diseases while all other tools were generic in nature.

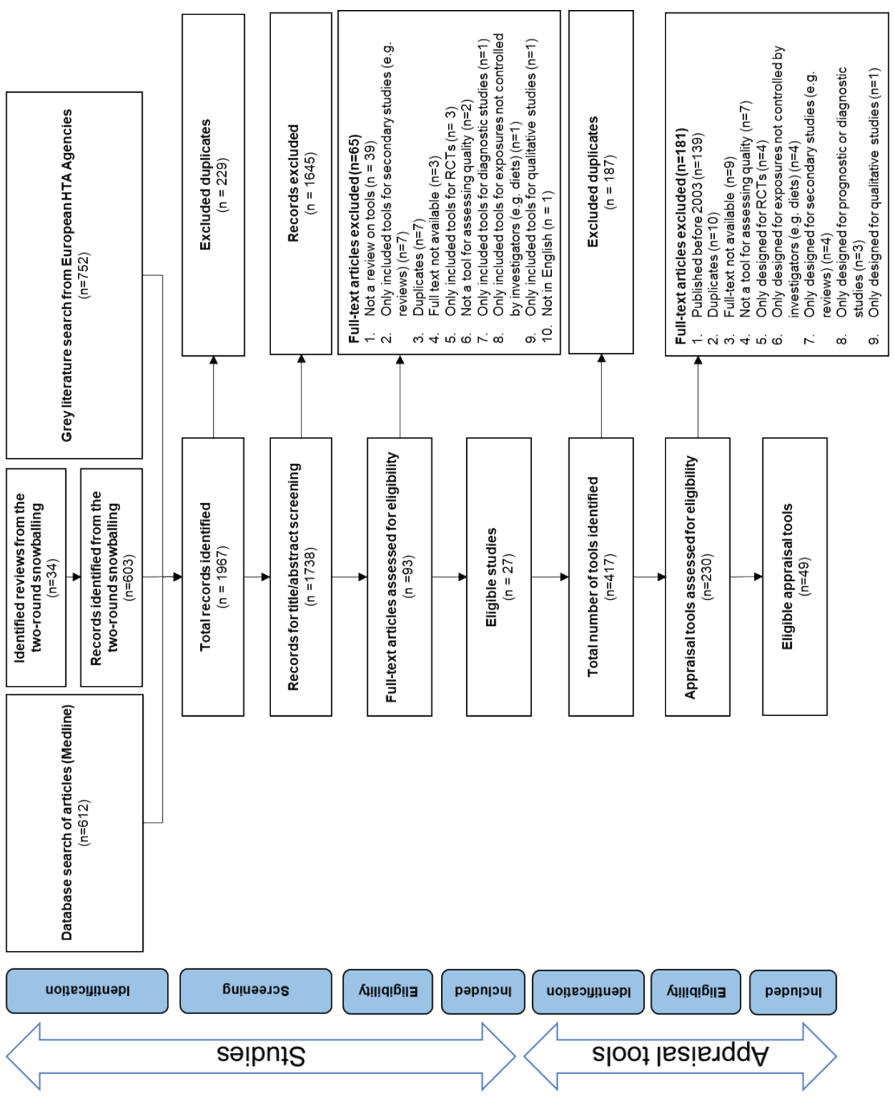



**Figure 1. Flow chart for the inclusion and exclusion of appraisal tools for non-randomized studies of interventions.**

**Table 1.** Characteristics of the 49 included appraisal tools for non-randomized studies of interventions

| Appraisal tools | Abbreviation | Publication year | Scope | Study design | Publication format |
|---|---|---|---|---|---|
| Critical Appraisal Skills Programme Tool | CASP | 2022 | M | Cohort, Case-control | GL, TD article |
| Joanna Briggs Institute's Critical Appraisal Tools | JBI | 2020 | M | Cohort, Case-control | GL, TD article |
| REal Life EVidence AssessmeNt Tool | RELEVANT | 2019 | M & R | All | TD article |
| STrengthening the Reporting of OBservational studies in Epidemiology Checklists | STROBE | 2019 | R | All | GL, TD article |
| Basque Office for Health Technology Assessment Tool | OSTEBA | 2019 | NR | All | GL |
| NA | Kennedy et al. 2019 | 2019 | R | All | TD article |
| Mixed Methods Appraisal Tool | MMAT | 2018 | M | All | TD article |
| Critical Appraisal Tools of Specialist Unit for Review Evidence | SURE | 2018 | M | Cohort, Cross-sectional, Case-control, Case-series | GL, TD article |
| NA | Viswanathan et al. 2018 | 2018 | M | All | TD article |
| European Network of Centres for Pharmacoepidemiology and Pharmacovigilance Guide on Methodological Standards in Pharmacoepidemiology | ENCePP | 2018 | R | All | GL, TD article |
| Risk Of Bias In Non-randomised Studies - of Interventions | ROBINS-I | 2017 | M & R | All | TD article |
| NA | Faillie et al. 2017 | 2017 | M | All | GL |
| Joint Task Force between the International Society for Pharmacoepidemiology and the International Society for Pharmacoeconomics and Outcomes Research | ISPE-ISPOR | 2017 | R | All | TD article |

**Table 1. Continued**

| Appraisal tools | Abbreviation | Publication year | Scope | Study design | Publication format |
|---|---|---|---|---|---|
| Appraisal tool for Cross-Sectional Studies | AXIS | 2016 | M | Cross-sectional | TD article |
| NA | Handu et al. 2016 | 2016 | M | All | TD article |
| International Society of Pharmacoepidemiology Guidelines for Good Pharmacoepidemiology Practice | ISPE | 2016 | NR | All | GL, TD article |
| REporting of studies Conducted using Observational Routinely-collected Data Checklist | RECORD | 2015 | M | All | TD article |
| Comparative Effectiveness Research Collaborative Initiative Questionnaire | CER-CI | 2014 | M | All | TD article |
| Good ReseArch for Comparative Effectiveness Checklist | GRACE | 2014 | NR | All | GL, TD article |
| Scottish Intercollegiate Guidelines Network Checklists | SIGN | 2014 | M | Cohort, Case-control | GL, TD article |
| A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions | ACROBAT-NRSI | 2014 | M | All | GL |
| Interventional Pain Management Techniques – Quality Appraisal of Reliability and Risk of Bias Assessment for Nonrandomized Studies | IPM-QRBNR | 2014 | M | All | TD article |
| Quality Assessment Tool of National Heart, Lung, and Blood Institute | NIH | 2013 | NR | Cohort, Cross-sectional, Case-control, Case-series | GL |
| Guidelines manual of National Institute for Health and Care Excellence: Appendices D-E | NICE | 2013 | M | Cohort, Case-control | GL |
| CAse REport (CARE) Guidelines Checklist | CARE | 2013 | M | Case-report | TD article |

**Table 1. Continued**

| Appraisal tools | Abbreviation | Publication year | Scope | Study design | Publication format |
|---|---|---|---|---|---|
| Institute of Health Economic Quality Appraisal Tool for Case-Series Studies | IHE | 2012 | M & R | Case-series | GL, TD article |
| Agency for Healthcare Research and Quality Methodology Checklist | AHRQ | 2012 | M | All | GL |
| Risk of Bias Assessment tool for Non-randomized Studies Tool | RoBANS | 2011 | M | All | GL, TD article |
| Research Triangle Institute Item Bank | RTI Item Bank | 2011 | M | All | GL, TD article |
| The Montreal Critical Appraisal Worksheet | Montreal | 2011 | R | All | GL |
| Grades of Recommendation, Assessment, Development and Evaluation Guideline | GRADE | 2011 | M & R | All | GL, TD article |
| NA | Blagojevic et al. 2010 | 2010 | M | Cohort, Case-control | Modifed for review |
| Academy of Nutrition and Dietetics Quality Criteria Checklist (Primary Research) | ANDQ | 2010 | M | All | GL |
| NA | Bishop et al. 2010 | 2009 | NR | All | Modifed for review |
| Harm Critical Appraisal Worksheet | Harm | 2009 | M | All | GL |
| Newcastle-Ottawa Scale | NOS | 2009 | NR | Cohort, Case-control | GL, TD article |
| NA | Pluye et al. 2009 | 2009 | M | All | TD article |
| NA | Young et al. 2009 | 2009 | M | All | TD article |
| NA | Atluri et al. 2008 | 2008 | M | All | Modifed for review |
| NA | Tseng et al. 2009 | 2008 | M | All | Modifed for review |
| NA | Heller et al. 2008 | 2008 | R | All | TD article |

**Table 1. Continued**

| Appraisal tools | Abbreviation | Publication year | Scope | Study design | Publication format |
|---|---|---|---|---|---|
| NA | Genaidy et al. 2007 | 2007 | M | All | TD article |
| Graphic Appraisal Tool for Epidemiological Studies | GATE | 2006 | M | All | TD article |
| NA | Weightman et al. 2004 | 2004 | M | All | GL |
| Transparent Reporting of Evaluations with Nonrandomized Designs | TREND | 2004 | M | All | GL, TD article |
| NA | Thomas et al. 2004 | 2004 | M | All | Modifed for review |
| NHS Wales Questions to Assist with the Critical Appraisal of a Cross-Sectional Study | NHS Wales | 2004 | M | Cross-sectional | GL |
| Methodology Index for Non-randomized Studies | MINORS | 2003 | M & R | All | TD article |
| NA | Rangel et al. 2003 | 2003 | NR | All | TD article |

NA indicates not applicable; NR, not reported; M, methodological quality; R, reporting; GL, grey literature; TD, tool development article; Modified for review: an appraisal tool modified from existing appraisal tools during a review of primary studies in a certain disease field or for a certain health intervention.

## Quality domains and items

We identified 44 criteria to describe study quality from three previous reviews (13-15). After merging criteria with similar content (e.g. "Follow-up" and "Loss to follow-up") and incorporating items into those with wider meanings (e.g. "Loss to follow-up bias" into "Loss to follow-up"), we obtained a list of 18 items. After the pilot coding, we summarized criteria of appraisal tools not covered by the 18 items into another eight items. According to the general order of conducting a NRSI (e.g. study design and data analysis, etc.), these 26 items were categorized into four domains: Study design, Data quality, Data analysis, and Results presentation. As shown in Figure 2 and Table 2, all domains and most items were addressed by existing tools, but for each item, the number of tools with sufficient descriptions was relatively small. For three items in methodology and for nine items in reporting, less than five tools addressed them, and none of the tools sufficiently described them.

Figure 2 illustrates whether and to what extent the identified tools addressed quality items in terms of methodological quality or reporting. The 26 columns represent the 26 quality items as shown in Table 2. The ranking of appraisal tools based on the number of items addressed or sufficiently described, either general or segmented by quality domains, is shown in Appendix 4-6. Regarding methodological quality, RTI Item Bank (27) addressed (n=20) and sufficiently described (n=18) the highest number of items. In addition, the tools that ranked both top 10, based on number of items addressed or sufficiently described, included MINORS (28), Faillie et al. 2017 (29), ROBINS-I (18), ANDQ (16), CER-CI (30), and JBI (31). These tools addressed at least 10 items, and sufficiently described at least 5 items. In the study-design domain, RTI Item Bank (27) and ROBINS-I (18) sufficiently described the most items (n>=5), while in the Data quality domain, RTI Item Bank (27), ANDQ (16), and MINOR (28) ranked the top three, which sufficiently described at least four of the 10 items. In the Data analysis domain, Faillie et al. 2017 (29) was the only tool that sufficiently described all the three included items, while the Mixed Methods Appraisal Tool (32), RTI Item Bank (27), and Viswanathan et al. 2018 (33) sufficiently described the two items. In the Results presentation domain, the relevant two items sufficiently described by Faillie et al. 2017 (29), Handu et al. 2016 (34), CER-CI (30), GRADE (35), and ANDQ (16). Regarding reporting, STROBE (36) addressed (n=17) and sufficiently described (n=13) the highest number of items. Also, the tools that ranked both top 10, based on the two criteria, included TREND (37), the tool by Genaidy et al. 2007 (38) , RECORD (39), ENCePP (40), ISPE (41), the tool by Tseng et al. 2009 (42), SURE (43), and ISPE-ISPOR (44). These tools at least addressed and sufficiently described seven and three quality items, respectively. In all the four quality domains, STROBE (34) sufficiently described the (equally) most items, compared to other tools. Besides, in the Study design domain, ENCePP (40), Genaidy et al. 2007 (38), and RECORD (39) sufficiently described at least four of the 11 items, while in the Data quality domain, RECORD (39), and TREND (37) sufficiently described at least four of the 10 items. In the Data analysis domain, STROBE was the only tool that sufficiently described two of the thee items, while 10 other tools, e.g., RELEVANT (45), sufficiently described only one item. In the Results presentation domain, Interventional Pain Management Techniques – Quality Appraisal of Reliability and Risk of Bias Assessment for Nonrandomized Studies (IPM-QRBNR) (46) sufficiently described all the two items.

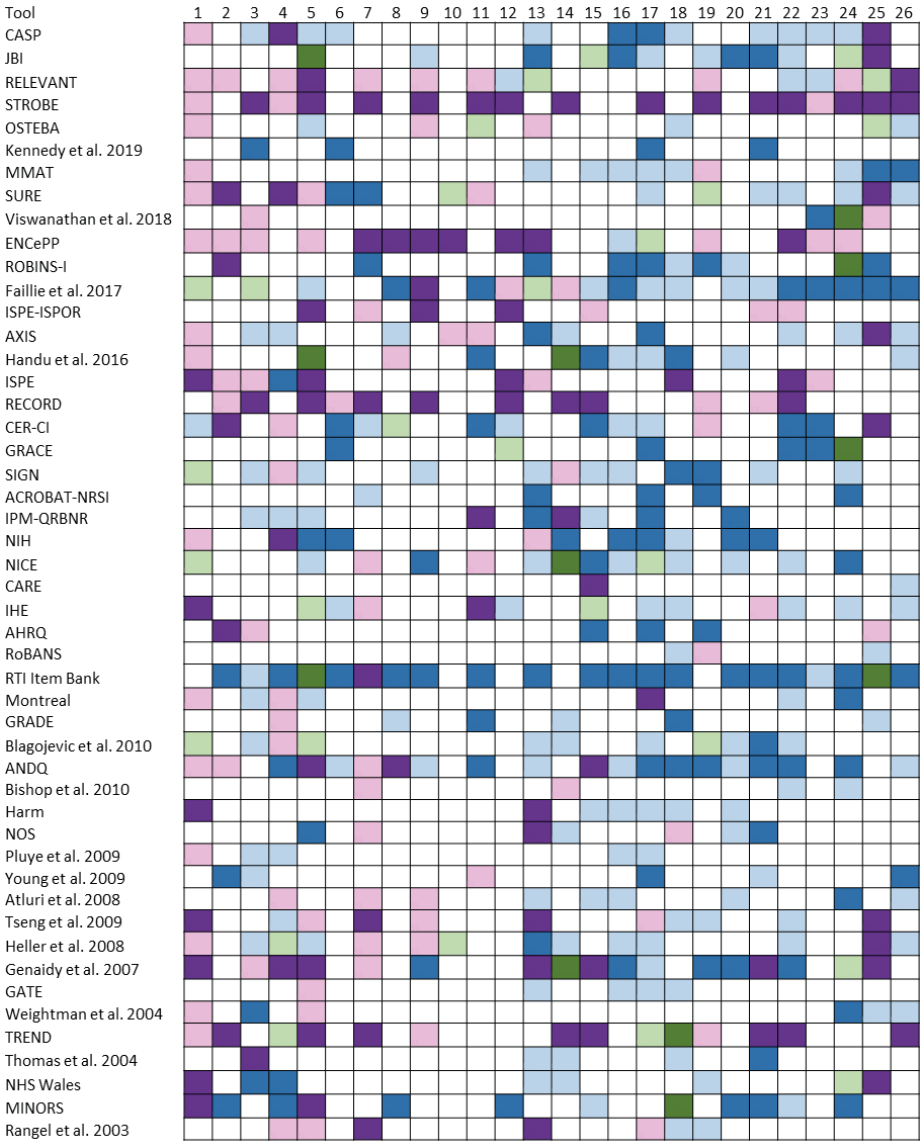



**Figure 2. The extent to which the appraisal tools addressed quality items on methodological quality or reporting.**

The 26 columns represent 26 quality items identified in the content analysis, and an overview of the items is shown in Table 2. Dark green indicates both methodology and reporting were sufficiently described in a tool; Light green, both methodology and reporting were addressed with only brief descriptions; Purple, reporting was addressed with sufficient descriptions; Pink, reporting was addressed with brief descriptions; Dark blue, methodological quality was addressed with sufficient descriptions; Light blue, methodological quality was addressed with brief descriptions.

**Table 2. Overview of the four domains and 26 quality items, with numbers and proportions of appraisal tools that addressed or sufficiently described them**

| Domains | Items | Number (%) of appraisal tools that addressed or sufficiently described a quality item | | | |
|---|---|---|---|---|---|
| | | Methodology (n=49) | | Reporting (n=49) | |
| | | Addressed | Sufficiently described | Addressed | Sufficiently described |
| 1. Study design | 1. Study objective | 6 (12) | 0 | 28 (57) | 7 (14) |
| | 2. Protocol | 5 (10) | 4 (8) | 8 (16) | 4 (8) |
| | 3. Selection of study design | 15 (31) | 3 (6) | 9 (18) | 3 (6) |
| | 4. Sample size/Power calculation | 11 (22) | 5 (10) | 15 (31) | 4 (8) |
| | 5. Eligibility criteria | 15 (31) | 5 (10) | 20 (41) | 12 (24) |
| | 6. Intervention selection | 9 (18) | 5 (10) | 2 (4) | 0 |
| | 7. Intervention definition | 4 (8) | 2 (4) | 17 (35) | 7 (14) |
| | 8. Outcome selection | 6 (12) | 3 (6) | 4 (8) | 2 (4) |
| | 9. Outcome definition | 7 (14) | 3 (6) | 11 (22) | 5 (10) |
| | 10. Ethical approval | 0 | 0 | 6 (12) | 2 (4) |
| | 11. Conflict of interest | 7 (14) | 6 (12) | 9 (18) | 3 (6) |
| 2. Data quality | 12. Data source | 5 (10) | 1 (2) | 7 (14) | 5 (10) |
| | 13. Patient recruitment | 19 (39) | 6 (12) | 12 (24) | 6 (12) |
| | 14. Participation rate | 12 (24) | 4 (8) | 10 (20) | 7 (14) |
| | 15. Baseline characteristics | 15 (31) | 5 (10) | 48 (16) | 5 (10) |
| | 16. Intervention measurement | 17 (35) | 6 (12) | 0 | 0 |
| | 17. Outcome measurement | 28 (57) | 12 (24) | 7 (14) | 2 (4) |
| | 18. Blinding of outcome | 22 (45) | 8 (16) | 4 (8) | 3 (6) |
| | 19. Missing data | 13 (27) | 7 (14) | 10 (20) | 1 (2) |
| | 20. Length of follow-up | 14 (29) | 5 (10) | 0 | 0 |
| | 21. Loss to follow-up | 15 (31) | 9 (18) | 6 (12) | 3 (6) |
| 3. Data analysis | 22. Description | 19 (39) | 6 (12) | 6 (12) | 5 (10) |
| | 23. Sensitivity analysis | 7 (14) | 3 (6) | 3 (6) | 0 |
| | 24. Bias adjustment | 22 (45) | 10 (20) | 5 (10) | 0 |
| 4. Results presentation | 25. Are all the results presented | 9 (18) | 4 (8) | 15 (31) | 11 (22) |
| | 26. Reaonable conclusions from results | 14 (29) | 4 (8) | 3 (6) | 3 (6) |

The judgement on whether criteria or signaling questions of an appraisal tool were relevant to methodological quality or reporting was made by authors, independently of what original studies claimed to be.

## Methodological quality

Among the four domains, the Study design domain was the most ignored domain by appraisal tools, as only four of the 11relevant items were described with sufficient details by more than four tools. More specifically, no tool described methodological quality on Ethical approval or Study objective with sufficient detail. For example, the guidelines manual of the National Institute for Health and Care Excellence (NICE) stated that: "The study addresses an appropriate and clearly focused question" (47). The tool did not explain the standard of appropriateness and clearness.

In addition, although one-third of tools discussed what a good study design was, only three tools defined the goodness (48-50). For example, the NHS Wales Questions to Assist with the Critical Appraisal of a Cross-Sectional Study (NHS Wales) stated that the choice of study design should be appropriate to the research question and ensure the reliability of study results (50). Outcome selection was also ignored by most tools, as only three tools (i.e. RTI Item Bank (27), MINORS (28), and the tool by Faillie et al. 2007 (29)) sufficiently described them. Similarly, only RTI Item Bank (27), the tool by Genaidy et al. 2007 (38), and NICE (45) sufficiently described the item Outcome definition. For example, Genaidy et al. 2007 stated that a definition was clear only if "definitions of all outcome variables were clearly described", and was partially clear if not all variables were clearly described, but "sufficient information was provided for the reader to understand the intent" (38).

Other items that were rarely addressed or insufficiently described included Intervention definition, Data source, and Sensitivity analysis. The respective tools with sufficient descriptions included SURE (43), ROBINS-I (18), MINORS (28), CER-CI (30), GRACE (17), and the tools described by Faillie et al. 2017 (29), and Viswanathan et al. 2018 (33).

## Reporting

The Data quality domain was ignored by most tools, as five of the 10 relevant items were sufficiently addressed by less than three tools. In particular, the item Intervention measurement and Length of follow-up were sufficiently addressed by none of the tools, JBI was the only tool stating that method of measuring interventions should be clearly reported (31), while 20 tools addressing Intervention measurement only focused on methodological quality. Some other items that were rarely addressed or insufficiently addressed included Outcome blinding and Loss to follow-up. Regarding Outcome blinding, only five tools provided sufficient descriptions, i.e., the tool by Faillie et al. 2017, RECORD, STROBE, ENCePP, and ISPOR-ISPE (29,39,41,40,44). Similarly, only the tool by Genaidy et al. 2007, TREND, and STROBE sufficiently described Loss to follow-up (38,39,34).

Another domain that was ignored was Data analysis. Only five tools, such as RELEVANT, emphasized the reporting of confounding (31,40,45,50). In addition, only three tools, i.e., STROBE (36), ISPE (41), and ENCePP (40), stated that sensitivity analyses should be reported. Still, a list of elements was lacking on what should be reported regarding confounding or sensitivity analyses.

## Discussion

We conducted a review of appraisal tools for non-randomized studies of interventions, and assessed whether and how sufficiently these tools addressed quality concerns, in terms of methodological quality or reporting, in four quality domains and across 26 items. Our study identified 49 tools, and showed that the RTI Item Bank and STROBE were most comprehensive, with the highest number of items addressed and sufficiently described, respectively, on methodological quality and reporting. However, none of the tools addressed concerns in all items, not even briefly. The items least addressed for methodological quality included outcome selection, outcome definition, and ethical approval, and for reporting included intervention selection, intervention measurement, and length of follow-up.



To our knowledge, this is the first study that compared level of sufficient descriptions of appraisal tools at quality-item levels. Previous reviews also compared appraisal tools but from different perspectives. D'Andrea et al. identified 44 tools evaluating the comparative safety and effectiveness of medications, and only assessed whether or not these tools addressed methodological quality in 8 domains (13). In another review, Lin-Lu Ma et al. elaborated for what types of study design a tool was suited (51).For example, for cohort studies, they encouraged using five tools, while discouraged the use of another two. However, they did not clarify why some tools were more suitable than the others. Quigley et al. identified 48 tools for appraising quality of systematic reviews of non-randomized studies, listed the five most commonly-used tools, and assessed whether they addressed the 12 quality domains, such as “appropriate design” and “appropriate statistical analysis” (14). Although the tools were compared using different criteria, some results were consistent among all studies. For example, both D'Andrea et al. (13) and our study found that intervention measurement, outcome measurement, and confounding were frequently addressed by existing tools. Also, Lin-Lu Ma et al. (51) and Quigley et al. (14) both recommended ROBINS-I, MINORS, and JBI, and all these tools ranked top 10 for addressing and sufficiently describing methodological quality in our study. With detailed information on level of sufficient descriptions of appraisal tools at the quality-item level, we add value to previous

reviews by listing quality concerns that such commonly recommended tools could not adequately address.

We also found some discrepancies in the tools identified or recommended. For example, of the 44 tools identified by D'Andrea et al. (13), 27 were published between 2003 and 2019; while in our study, 47 were identified as published between 2003 and 2019. This discrepancy could be explained by additional tools identified through other reviews, tools from grey literature, and differences in eligibility criteria (e.g. exclusion of non-pharmacological interventions or assessing only one or a few specific types of bias). Another discrepancy was that some tools that ranked top in our study were less recommended by previous reviews, such as RTI Item Bank (27) and the tool by Faillie et al. 2017 (29) for methodological quality and by Genaidy et al. 2007 (38) for reporting. This might be explained by the novel criteria (i.e. how sufficiently quality items were addressed) we used to evaluate these tools.

We discovered that, with information on how sufficiently a tool described a quality item, tool users might broaden their horizons on quality concerns of non-randomized studies to be considered. For example, if ROBINS-I (18) is used for assessing methodological quality, the quality concerns known to users will be RoB in eight domains (e.g. confounding and selection bias). However, as shown in Figure 2, quality concerns in 16 items (e.g. Intervention selection and Outcome definition) may not be sufficiently described in ROBINS-I but in other tools, such as RTI item bank (27), the NICE checklist (47), and the tool by NHS Wales (50). Similarly, if users check the ENCePP (40) and ISPE tools (41), in addition to STROBE, for reporting quality concerns, they may more comprehensively understand concerns on Ethical approval, Outcome definition, Study objective, and Data source. Tool users who may benefit from such information are not only researchers who conduct non-randomized studies and decision-makers who assess study quality, but also tool developers who may identify a research gap.

While the needs of tool users may vary, they could all be somewhat satisfied by our research. For example, it is important for researchers to ensure sufficient reporting of the strengths and weaknesses of a NRSI, as such information will be ultimately used for determining the eligibility of their studies for a decision-making (34,52). For HTA agencies, NRSIs can be used to extrapolate long-term drug effectiveness and to identify drug-related costs, and a deep and consistent understanding of how to assess NRSI quality among the agencies is important for promoting the use of real-world data (53). For regulators, a comprehensive understanding of how to evaluate NRSI quality may promote a structured pattern of using RWD to support drug regulation (54). While

researchers focus more on reporting, and decision-makers (e.g. HTA agencies) have emphasis on methodological quality, we suggest all users pay attention to the linkage between methodology and reporting for each quality item, as illustrated in our research, as it could help understand the necessity of investigating each item.

Another finding of our research was that whether and to what extent a quality concern was addressed by a tool partly depended on the tool purpose. For example, the GRACE checklist was designed as a "screening tool" to exclude studies that did not meet basic quality requirements (17), and ROBINS-I focused on RoB, rather than all methodological quality issues, such as appropriateness of study objectives or statistical analyses for patient matching (18). Some tools, such as JBI Cohort (31), were specific to a type of study design. While they addressed less than half of quality items defined in our research, they were proven robust in many studies (14). Additionally, for several quality items we found some heterogeneity in content of signaling questions or criteria among the tools with sufficient description. For example, to assess methodological quality of sensitivity analysis, CER-CI (30) stated that key assumptions or definitions of outcomes should be tested, while the tool by Viswanathan et al. 2018 (33) emphasized the importance of reducing uncertainty in individual judgements. Given the heterogeneity of tools, we suggest users following a two-step approach when selecting a tool. First, users may narrow down the scope of tools based on their own needs, e.g. excluding tools for a different study design. This step could be achieved by referring to synthesized results and recommendations from existing reviews (13,14). Second, users could use the overview we provide (Figure 2) to see which tool(s) could provide complementary insights the tool of their first choice is lacking.



Furthermore, we found that appraisal tools designed for specific interventions had potential to be transferred for general interventions. In our research, the tools described by Tseng et al. 2009 (42), and Blagojevic et al. 2010 (55), and ANDQ (16), were originally designed for a surgical intervention, knee osteoarthritis, and for the field of diabetes, respectively. All these tools ranked top 15 in our study for addressing either methodological quality or reporting (Appendix 4-6), and many of their criteria could be generalizable. For example, Tseng et al. 2009 stated that interventions could be adequately described with specifically referenced articles (42). Though such tools could be transferred, they often used disease-or-intervention-specific concepts in their criteria, which might be adjusted before being applied more widely.

Moreover, we noticed that, some quality items were less frequently addressed, such as Study objective, Ethical approval, or Sensitivity analysis, compared to other items. This might be explained by the fact that, some items were more related to a certain

purpose of tool application than the others. For example, a tool addressing concerns on RoB may focus less on Study objective, which is relatively more difficult to be directly linked to a well-defined type of bias. Still, since these quality items are related to NRSI quality, and they are rarely sufficiently described, particular efforts investigating these quality items may be needed in future tool development. Also, tools designed for a specific purpose may make users realize that some items are beyond their scope, but still need to be paid attention.

Our study has a number of limitations. One limitation is that, some tools identified by our study were originally developed for purposes beyond assessing methodological quality of reporting of NRSIs, so our study could not cover all potentials of these tools. For example, the GRADE framework was mainly designed for addressing certainty of evidence, such as indirectness (i.e. whether interventions were compared directly), and for making relevant clinical practice recommendations. While it mentions RoB (e.g. publication bias), its main purpose is to illustrate how to grade quality of evidence, rather than to function as an exact quality appraisal tool. In other words, the GRADE allows users to use any additional tools to assess NRSI quality (35). Also, the GRADE checklist was designed for both RCTs and NRSIs, so some criteria might be relatively brief, compared to specifically-designed tools, such as RTI Item Bank (27). Finally, GRADE can be used to estimate and score the quality of evidence for the full body of evidence and not only for individual primary studies. Therefore, tool users who assess NRSIs beyond methodological quality or reporting should consider criteria in addition to those mentioned in our study, for selecting a tool. Another limitation is that, some tools were predecessors of others, but we did not exclude them if they met the inclusion criteria. For example, the ROBINS-I tool was developed from the Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI) (56), and some of their signaling questions differed. Such information on tool linkage may also be considered for tool selection, if available from the tools. Another limitation is that we only searched HTA agencies for grey literature, and the returned hits on the snowballing approach depended on the starting-set articles, so some tools only mentioned by clinical guideline or regulatory organizations, or tools missed by the previous reviews might have been overlooked. Also, only one researcher (MH) traced versions of tools, by following reference lists of the identified studies and by visiting websites of the online tools. Consequently, the most up-to-date version of a tool might be missing, and the extent to which a quality item was described by a tool might be underestimated. As appraisal tools are developed or improved continuously, an online platform that automatically identifies appraisal tools and summarizes tool information is promising. Such platforms have already been established for tools for assessing observational studies for exposures that were not controlled by investigators

(e.g. dietary patterns) (57). Another limitation is that we categorized criteria of a quality item as "sufficient" or "brief" for each tool, based on whether an explanation was provided for the criteria. Though consensus was reached among authors, and all tool criteria were independently reviewed by two researchers, tool users might question the feasibility of such categorization when selecting a tool. Hence, further case studies with expert inputs may be needed to test whether a tool selected based on such categorization, together with recommendations from previous reviews, can really satisfy tool users. It is also worth noting that, the target audience of this review and content analysis could be decision-makers who assess the general quality of a NRSI, NRSI performers who may report quality of their studies, or developers of relevant appraisal tools. However, when users focus on a specific type of concern (e.g. causal effect or data quality), some methodological guidance investigating the specific issue or tools beyond the healthcare field (e.g. social science) really exist (58,59), and might include signaling questions that address additional concerns. These literature should be referred to by users. Moreover, the tools for diagnosis studies, prognosis studies, and secondary studies were beyond the scope of our study, and relevant users may refer to other studies, such as Quigley et al. (2019) (14), for further information.



## Conclusions

Most of the appraisal tools for non-randomized studies of interventions have their own strengths, but none of them could address all quality concerns relevant to these studies. Even the most comprehensive tools could be complemented with items from other tools. With information on how sufficiently a tool describes a quality item, tool users might broaden their horizons on quality concerns of non-randomized studies to be considered, and might select a tool that more completely satisfies their needs. We suggest decision-makers, researchers, and tool developers consider the quality-item level heterogeneity when selecting a tool or identifying a research gap.

## Author contribution

LJ designed the study protocol, identified appraisal tools, conducted the content analysis, and wrote the manuscript. All co-authors contributed to this study and critically reviewed and approved the manuscript.

# References


1. Makady A, de Boer A, Hillege H, et al. What is real-world data? A review of definitions based on literature and stakeholder interviews. Value Health. 2017 Jul 1;20(7):858-65.
2. Reeves BC, Deeks JJ, Higgins JPT, Shea B, Tugwell P, Wells GA. Including non-randomized studies on intervention effects. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022) (Internet).Cochrane;2022.
3. ROBINS-E Development Group (Higgins J, Morgan R, Rooney A, Taylor K, Thayer K, Silva R, Lemeris C, Akl A, Arroyave W, Bateson T, Berkman N, Demers P, Forastiere F, Glenn B, Hróbjartsson A, Kirrane E, LaKind J, Luben T, Lunn R, McAleenan A, McGuinness L, Meerpohl J, Mehta S, Nachman R, Obbagy J, O'Connor A, Radke E, Savović J, Schubauer-Berigan M, Schwingl P, Schunemann H, Shea B, Steenland K, Stewart T, Straif K, Tilling K, Verbeek V, Vermeulen R, Viswanathan M, Zahm S, Sterne J). (Internet) Risk Of Bias In Non-randomized Studies - of Exposure (ROBINS-E). 2022.
4. Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. Multidiscip Healthc. 2018 Jul 2:295-304. doi: 10.2147/JMDH.S160029. eCollection 2018.
5. Baumfeld Andre E, Carrington N, Siami FS, et al. The Current Landscape and Emerging Applications for Real-World Data in Diagnostics and Clinical Decision Support and its Impact on Regulatory Decision Making. Clin Pharmacol Ther. 2022 Dec;112(6):1172-82.
6. Makady A, van Veelen A, Jonsson P, et al. Using real-world data in health technology assessment (HTA) practice: a comparative study of five HTA agencies. Pharmacoeconomics. 2018 Mar;36:359-68.
7. Kent S, Salcher-Konrad M, Boccia S, et al. The use of nonrandomized evidence to estimate treatment effects in health technology assessment. J Comp Eff Res. 2021 Jun;10(14):1035-43.
8. Facey KM, Rannanheimo P, Batchelor L, et al. Real-world evidence to support Payer/HTA decisions about highly innovative technologies in the EU—actions for stakeholders. Int J Technol Assess Health Care. 2020 Aug;36(4):459-68.
9. Hogervorst MA, Pontén J, Vreman RA, et al. Real world data in health technology assessment of complex health technologies. Front Pharmacol. 2022:297.
10. Schünemann HJ, Tugwell P, Reeves BC, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. Res Synth Methods. 2013 Mar;4(1):49-62.
11. Jonathan AC Sterne, Miguel A Hernán, Alexandra McAleenan, Barnaby C Reeves, Julian PT Higgins. Assessing risk of bias in a non-randomized study. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022) .Cochrane;2022.
12. Orsini LS, Berger M, Crown W, et al. Improving transparency to build trust in real-world secondary data studies for hypothesis testing—why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. Value Health. 2020 Sep 1;23(9):1128-36.
13. D'Andrea E, Vinals L, Patorno E, et al. How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools. BMJ Open. 2021 Mar 1;11(3):e043961.
14. Quigley JM, Thompson JC, Halfpenny NJ, et al. Critical appraisal of nonrandomized studies—a review of recommended and commonly used tools. J Eval Clin Pract. 2019 Feb;25(1):44-52.
15. University of Sheffield. The use of observational data to inform estimates of treatment effectiveness in technology appraisal: Methods for comparative individual patient data. 2015 May. Available from : https://www.sheffield.ac.uk/nice-dsu/tsds/full-list . [Accessed Feb 8, 2023].

16. Evidence Analysis Library (EAL). Quality Criteria Checklist: Primary Research. Available from: https://www.andeal.org/vault/2440/web/files/QCC_3.pdf. [Accessed Feb 8, 2023].
17. Dreyer NA, Velentgas P, Westrich K et al. The GRACE checklist for rating the quality of observational studies of comparative effectiveness: a tale of hope and caution. J Manag Care Pharm. 2014 Mar;20(3):301-8.
18. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016 Oct 12;355.
19. HTx. About HTx project. Available from: https://www.htx-h2020.eu/about-htx-project. [Accessed Oct 25, 2022].
20. Open Science Framework (OSF) Registry. About. Available from: https://osf.io/dashboard. [Accessed Oct 25, 2022].
21. Wohlin C. Second-generation systematic literature studies using snowballing. InProceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering 2016 Jun 1 (pp. 1-6).
22. Connected Papers. About. Available from : https://www.connectedpapers.com/about. [Accessed Oct 25, 2022].
23. The International Network of Agencies for Health Technology Assessment . Members. Available from: https://www.inahta.org/members/members_list. [Accessed Oct 25, 2022].
24. Ouzzani M, Hammady H, Fedorowicz Z, et al. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016 Dec;5:1-0.
25. Whiting P, Wolff R, Mallett S, et al. A proposed framework for developing quality assessment tools. Syst Rev. 2017 Dec;6:1-9.
26. Nowell LS, Norris JM, White DE, et al. Thematic analysis: Striving to meet the trustworthiness criteria. Int J Qual Methods. 2017 Sep 28;16(1):1609406917733847.
27. Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. J Clin Epidemiol. 2012 Feb 1;65(2):163-78. doi: 10.1016/j.jclinepi.2011.05.008.
28. Slim K, Nini E, Forestier D, et al. Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. ANZ J Surg. 2003 Sep;73(9):712-6.
29. Faillie JL, Ferrer P, Gouverneur A et al. A new risk of bias checklist applicable to randomized trials, observational studies, and systematic reviews was developed and validated to be used for systematic reviews focusing on drug adverse events. J Clin Epidemiol. 2017 Jun 1;86:168-75.
30. Berger ML, Martin BC, Husereau D, et al. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health. 2014 Mar 1;17(2):143-56
31. Joanna Briggs Institute. Critical Appraisal Tools. Available from: https://jbi.global/critical-appraisal-tools. [Accessed Feb 8, 2023].
32. Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). J Eval Clin Pract. 2018 Jun;24(3):459-67.
33. Viswanathan M, Patnode CD, Berkman ND, et al. Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. J Clin Epidemiol. 2018 May 1;97:26-34.
34. Handu D, Moloney L, Wolfram T, et al. Academy of Nutrition and Dietetics methodology for conducting systematic reviews for the Evidence Analysis Library. J Acad Nutr Diet. 2016 Feb;116(2):311-8.

35. Guyatt GH, Oxman AD, Vist G, GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J. Clin. Epidemiol. 2011 Apr 1;64(4):407-15.

36. Vandenbroucke JP, Elm EV, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Ann Intern Med. 2007 Oct 16;147(8):W-163.

37. Des Jarlais DC, Lyles C, Crepaz N, Trend Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health. 2004 Mar;94(3):361-6.

38. Genaidy AM, Lemasters GK, Lockey J, et al. An epidemiological appraisal instrument–a tool for evaluation of epidemiological studies. Ergonomics. 2007 Jun 1;50(6):920-60.

39. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS medicine.

40. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance Guide on Methodological Standards in Pharmacoepidemiology (ENCePP) (Internet). ENCePP Guide on Methodological Standards in Pharmacoepidemiology (Revision 10), 2022 June 30.Available from: https://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml. [Accessed Feb 8, 2023].

41. Public Policy Committee, International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practice (GPP). Pharmacoepidemiol Drug Saf. 2016 Jan;25(1):2-10.

42. Tseng TY, Breau RH, Fesperman SF, et al. Evaluating the evidence: the methodological and reporting quality of comparative observational studies of surgical interventions in urological publications. BJU Int. 2009 Apr;103(8):1026-31.

43. Cardiff University. Critical appraisal tools. Available from:https://www.cardiff.ac.uk/specialist-unit-for-review-evidence/resources/critical-appraisal-checklists. [Accessed Oct 25, 2022].

44. Wang SV, Schneeweiss S, Berger ML, et al. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1. 0. Value Health. 2017 Sep 1;20(8):1009-22.

45. Campbell JD, Perry R, Papadopoulos NG, et al. The REal Life EVidence AssessmeNt Tool (RELEVANT): development of a novel quality assurance asset to rate observational comparative effectiveness research studies. Clin Transl Allergy. 2019;9(1):21.

46. Manchikanti L, Hirsch JA, Heavner JE, et al. Development of an interventional pain management specific instrument for methodologic quality assessment of nonrandomized studies of interventional techniques. Pain Physician. 2014;17(3):E291.

47. National Institute for Health and Care Excellence (NICE). The guidelines manual:

48. Kennedy CE, Fonner VA, Armstrong KA, et al. The Evidence Project risk of bias tool: assessing study rigor for both randomized and non-randomized intervention studies. Syst Rev. 2019 Dec;8(1):1-0.

49. NHS Wales. A systematic approach to identifying the evidence. Project methodology 5. Cardiff: Information Services UWCM; 2004 Jan. Available from: https://www2.nphs.wales.nhs.uk/VulnerableAdultsDocs.nsf/0/3811E6F969F2D3FC8025783E005B59AB/$file/housingrelatedsupport_descriptive_evidencereview_final_200111.doc?OpenElement. [Accessed 2022 Oct 25].

50. NHS Wales. Questions to assist with the critical appraisal of a cross-sectional study (Type IV evidence). Available from: https://www2.nphs.wales.nhs.uk/PubHObservatoryProjDocs.nsf/($All)/E7B0C80995DC1BA380257DB80037C699/$File/Cross%20sectional%20study%20checklist.docx?OpenElement. [Accessed 2022 Oct 25]

51. Ma LL, Wang YY, Yang ZH, et al. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better?. Mil Med Res. 2020 Dec;7:1-1.

52. Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ. 2001;323:42–6.

53. Makady A, van Veelen A, Jonsson P, Moseley O, D'Andon A, de Boer A, Hillege H, Klungel O, Goettsch W. Using real-world data in health technology assessment (HTA) practice: a comparative study of five HTA agencies. Pharmacoeconomics. 2018 Mar;36:359-68.

54. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. Clin Pharmacol Ther. 2019 Apr;105(4):867-77.

55. Blagojevic M, Jinks C, Jeffery A, et al. Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis. Osteoarthritis Cartilage. 2010 Jan 1;18(1):24-33.

56. University of Bristol. Archived tool: A Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions (ACROBAT-NRSI). Available from: https://www.bristol.ac.uk/population-health-sciences/centres/cresyda/barr/riskofbias/robins-i/acrobat-nrsi (Accessed Oct 10, 2023)

57. Wang Z, Taylor K, Allman-Farinelli M, et al. A systematic review: Tools for assessing methodological quality of human observational studies. MedRxiv. (Preprint). May 21, 2019. https://doi.org/10.31222/osf.io/pnqmy

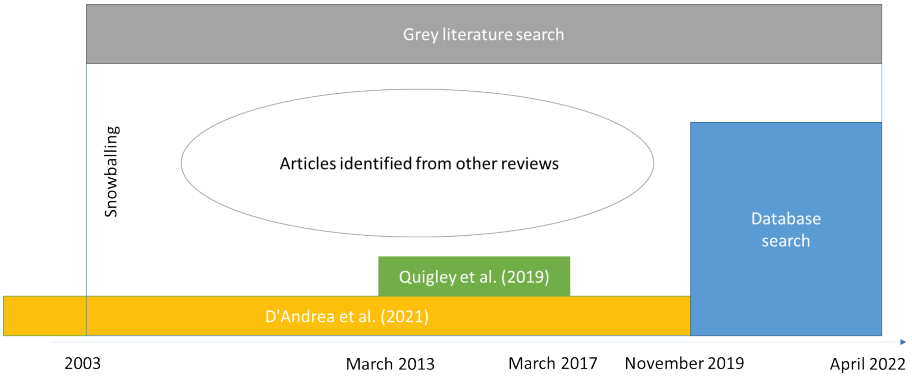
58. Jaksa A, Wu J, Jónsson P, Eichler HG, Vititoe S, Gatto NM. Organized structure of real-world evidence best practices: moving from fragmented recommendations to comprehensive guidance. J Comp Eff Res. 2021 Jun;10(9):711-31.

59. WILEY online library. How to Appraise the Studies: An Introduction to Assessing Study Quality. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9780470754887.ch5. [Accessed Oct 10, 2023]

60. Sharma Waddington H, Cairncross S. PROTOCOL: Water, sanitation and hygiene for reducing childhood mortality in low-and middle-income countries. Campbell Syst Rev. 2021 Mar;17(1):e1135.

# Appendices

**Appendix 1. A diagram illustrating how the three literature review approaches complemented each other**




**Appendix 2. Reference list of the included studies reviewing or describing appraisal tools for non-randomized studies of interventions**

1. D'Andrea E, Vinals L, Patorno E, Franklin JM, Bennett D, Largent JA, Moga DC, Yuan H, Wen X, Zullo AR, Debray TP. How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools. BMJ open. 2021 Mar 1;11(3):e043961.
2. Sanderson S, Tatt ID, Higgins J. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. International journal of epidemiology. 2007 Jun 1;36(3):666-76.Methodological quality assessment tools of non-experimental studies: a systematic review evaluating non-randomised intervention studies.
3. Scott HS. Systems to rate the strength of scientific evidence.
4. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. International Journal for Quality in Health Care. 2004 Feb 1;16(1):9-18.
5. Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng XT. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better?. Military Medical Research. 2020 Dec;7:1-1.
6. Zeng X, Zhang Y, Kwong JS, Zhang C, Li S, Sun F, Niu Y, Du L. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. Journal of evidence-based medicine. 2015 Feb;8(1):2-10.
7. Farrah K, Young K, Tunis MC, Zhao L. Risk of bias tools in systematic reviews of health interventions: an analysis of PROSPERO-registered protocols. Systematic reviews. 2019 Dec;8(1):1-9.
8. Quigley JM, Thompson JC, Halfpenny NJ, Scott DA. Critical appraisal of nonrandomized studies—a review of recommended and commonly used tools. Journal of Evaluation in Clinical Practice. 2019 Feb;25(1):44-52.
9. Patole S. Systematic Reviews and Meta-Analyses of Non-randomised Studies. InPrinciples and Practice of Systematic Reviews and Meta-Analysis 2021 Jun 27 (pp. 139-146). Cham: Springer International Publishing.

10. Zeng X, Zhang Y, Kwong JS, Zhang C, Li S, Sun F, Niu Y, Du L. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. Journal of evidence-based medicine. 2015 Feb;8(1):2-10.

11. Page MJ, McKenzie JE, Higgins JP. Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review. BMJ open. 2018 Mar 1;8(3):e019703.

12. Waddington H, Aloe AM, Becker BJ, Djimeu EW, Hombrados JG, Tugwell P, NOS G, Reeves B. Quasi-experimental study designs series—paper 6: risk of bias assessment. Journal of Clinical Epidemiology. 2017 Sep 1;89:43-52.

13. Saunders LD, Soomro GM, Buckingham J, Jamtvedt G, Raina P. Assessing the methodological quality of nonrandomized intervention studies. Western Journal of Nursing Research. 2003 Mar;25(2):223-37.

14. Brand J, Hardy R, Monroe E. Research pearls: Checklists and flowcharts to improve research quality. Arthroscopy: The Journal of Arthroscopic & Related Surgery. 2020 Jul 1;36(7):2030-8.

15. Lundh A, Rasmussen K, Østengaard L, Boutron I, Stewart LA, Hróbjartsson A. Systematic review finds that appraisal tools for medical research studies address conflicts of interest superficially. Journal of Clinical Epidemiology. 2020 Apr 1;120:104-15.

16. Yao X, Florez ID, Zhang P, Zhang C, Zhang Y, Wang C, Liu X, Nie X, Wei B, Ghert MA. Clinical research methods for treatment, diagnosis, prognosis, etiology, screening, and prevention: a narrative review. Journal of Evidence-Based Medicine. 2020 May;13(2):130-6.

17. Liebherz S, Schmidt N, Rabung S. How to assess the quality of psychotherapy outcome studies: A systematic review of quality assessment criteria. Psychotherapy Research. 2016 Sep 2;26(5):573-89.

18. Tate RL, Douglas J. Use of reporting guidelines in scientific writing: PRISMA, CONSORT, STROBE, STARD and other resources. Brain Impairment. 2011 May;12(1):1-21.

19. Viswanathan et al. 2018 M, Berkman ND, Dryden DM, Hartling L. Assessing risk of bias and confounding in observational studies of interventions or exposures: further development of the RTI item bank.

20. Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, Davies P, Kleijnen J, Churchill R. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. Journal of clinical epidemiology. 2016 Jan 1;69:225-34.

21. Public health department in Saint-Denis. Real-world studies for the assessment of medicinal products and medical devices. Available from : https://www.has-sante.fr/upload/docs/application/pdf/2021-06/real-world_studies_for_the_assessment_of_medicinal_products_and_medical_devices.pdf. [Accessed Dec 6, 2023].

22. Institute of Health Economics. Quality Appraisal Checklist for Case Series Studies and Instructions for Use. Available from: https://www.ihe.ca/publications/ihe-quality-appraisal-checklist-for-case-series-studies. [Accessed Dec 6, 2023].

23. University of Alberta: Education and Research Archive. Standard quality assessment criteria for evaluating primary research papers from a variety of fields. Available from: https://era.library.ualberta.ca/items/48b9b989-c221-4df6-9e35-af782082280e. [Accessed Dec 6, 2023].

24. National Institute for Health and Care Excellence. NICE real-world evidence framework. Available from : https://www.nice.org.uk/corporate/ecd9/chapter/overview. [Accessed Dec 6, 2023]

25. Lewin S, Booth A, Glenton C, Munthe-Kaas H, Rashidian A, Wainwright M, Bohren MA, Tunçalp Ö, Colvin CJ, Garside R, Carlsen B. Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. Implementation Science. 2018 Jan;13(1):1-0.

26. Institute of Health Economics. Development of a quality appraisal tool for case series studies using a modified Delphi technique. Available from: https://cobe.paginas.ufsc.br/files/2014/10/MOGA.Case-series.pdf. [Accessed Dec 6, 2023].

27. Agency for Healthcare Research and Quality. A Guide to Real-World Evaluations of Primary Care Interventions: Some Practical Advice. Available from: https://www.ahrq.gov/ncepcr/tools/pcmh/implement/evaluation-guide.html. [Accessed Dec 6, 2023].

**Appendix 3. Reference list of appraisal tools for non-randomized studies of interventions**

| Full name | Abbreviation | Reference |
|---|---|---|
| REal Life EVidence AssessmeNt Tool | RELEVANT | Campbell JD, Perry R, Papadopoulos NG, Krishnan J, Brusselle G, Chisholm A, Bjermer L, Thomas M, Van Ganse E, Van Den Berge M, Quint J. The REal Life EVidence AssessmeNt Tool (RELEVANT): development of a novel quality assurance asset to rate observational comparative effectiveness research studies. Clinical and translational allergy. 2019;9(1):21. |
| Graphic Appraisal Tool for Epidemiological Studies | GATE | Jackson R, Ameratunga S, Broad J, Connor J, Lethaby A, Robb G, NOS S, Glasziou P, Heneghan C. The GATE frame: critical appraisal with pictures. BMJ Evidence-Based Medicine. 2006 Apr 1;11(2):35-8. |
| Mixed Methods Appraisal Tool | MMAT | Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). Journal of evaluation in clinical practice. 2018 Jun;24(3):459-67. |
| Critical Appraisal Skills Programme Tool | CASP | Long, H. A., French, D. P., & Brooks, J. M. (2020). Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Research Methods in Medicine & Health Sciences, 1(1), 31-42. |
| Critical Appraisal Tools of Specialist Unit for Review Evidence | SURE | Available from : https://www.cardiff.ac.uk/specialist-unit-for-review-evidence/resources/critical-appraisal-checklists. |
| Joanna Briggs Institute's Critical Appraisal Tools | JBI | Available from : https://jbi.global/critical-appraisal-tools. |
| Risk Of Bias In Non-randomised Studies - of Interventions | ROBINS-I | Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. bmj. 2016 Oct 12;355. |
| Comparative Effectiveness Research Collaborative Initiative Questionnaire | CER-CI | Berger ML, Martin BC, Husereau D, Worley K, Allen JD, Yang W, Quon NC, Mullins CD, Kahler KH, Crown W. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value in health. 2014 Mar 1;17(2):143-56. |

**Appendix 3. Continued**

| Full name | Abbreviation | Reference |
|---|---|---|
| Good ReseArch for Comparative Effectiveness Checklist | GRACE | Dreyer NA, Velentgas P, Westrich K, Dubois R. The GRACE checklist for rating the quality of observational studies of comparative effectiveness: a tale of hope and caution. Journal of Managed Care Pharmacy. 2014 Mar;20(3):301-8. |
| Quality Assessment Tool of National Heart, Lung, and Blood Institute | NIH | Available from : https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools. |
| NA | Weightman et al. 2004 | Weightman AL, Mann MK, Sander L, Turley RL. Health evidence bulletins Wales: A systematic approach to identifying the evidence. Project methodology 5. Cardiff: Information Services UWCM. 2004. |
| Risk of Bias Assessment tool for Non-randomized Studies Tool | RoBANS | Available from : https://abstracts.cochrane.org/2011-madrid/risk-bias-assessment-tool-non-randomized-studies-robans-development-and-validation-new. |
| Research Triangle Institute Item Bank | RTI Item Bank | Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. Journal of clinical epidemiology. 2012 Feb 1;65(2):163-78. |
| Scottish Intercollegiate Guidelines Network Checklists | SIGN | Available from : https://www.sign.ac.uk/what-we-do/methodology/checklists. |
| The Montreal Critical Appraisal Worksheet | Montreal | Available from : https://guides.bib.umontreal.ca/ckfinder/ckeditor_assets/attachments/critical-appraisal-worksheet.pdf. |
| STrengthening the Reporting of OBservational studies in Epidemiology Checklists | STROBE | Available from : https://www.strobe-statement.org. |
| Transparent Reporting of Evaluations with Nonrandomized Designs | TREND | Des Jarlais DC, Lyles C, Crepaz N, Trend Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. American journal of public health. 2004 Mar;94(3):361-6. |
| A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions | ACROBAT-NRSI | Sterne JA, Higgins J, Reeves B. A Cochrane risk of bias assessment tool: for non-randomized studies of interventions (ACROBAT-NRSI). Version. 2014 Sep;1(0):24. |
| Methodology Index for Non-randomized Studies | MINORS | Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. ANZ journal of surgery. 2003 Sep;73(9):712-6. |

**Appendix 3. Continued**

| Full name | Abbreviation | Reference |
|---|---|---|
| Grades of Recommendation, Assessment, Development and Evaluation Guideline | GRADE | Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). Journal of clinical epidemiology. 2011 Apr 1;64(4):407-15. |
| NA | Rangel et al. 2003 | Rangel SJ, Kelsey J, Colby CE, Anderson J, Moss RL. Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. Journal of pediatric surgery. 2003 Mar 1;38(3):390-6. |
| NA | Thomas et al. 2004 | Thomas BH, Ciliska D, Dobbins M, Micucci S. A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. Worldviews on Evidence-Based Nursing. 2004 Sep;1(3):176-84. |
| NA | Atluri et al. 2008 | Atluri S, Datta S, Falco F, Lee M. Systematic review of diagnostic utility and therapeutic effectiveness of thoracic facet joint interventions. Pain Physician. 2008;11(5):611. |
| NA | Bishop et al. 2010 | Bishop FL, Prescott P, Chan YK, Saville J, von Elm E, Lewith GT. Prevalence of complementary medicine use in pediatric cancer: a systematic review. Pediatrics. 2010 Apr;125(4):768-76. |
| NA | Blagojevic et al. 2010 | Blagojevic M, Jinks C, Jeffery A, Jordan 1. Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis. Osteoarthritis and cartilage. 2010 Jan 1;18(1):24-33. |
| NA | Genaidy et al. 2007 | Genaidy AM, Lemasters GK, Lockey J, Succop P, Deddens J, Sobeih T, Dunning K. An epidemiological appraisal instrument–a tool for evaluation of epidemiological studies. Ergonomics. 2007 Jun 1;50(6):920-60. |
| Harm Critical Appraisal Worksheet | Harm | Available from : https://www.colleaga.org/tools/harm-critical-appraisal-worksheet. |
| NA | Tseng et al. 2009 | Tseng TY, Breau RH, Fesperman SF, Vieweg J, Dahm P. Evaluating the evidence: the methodological and reporting quality of comparative observational studies of surgical interventions in urological publications. BJU international. 2009 Apr;103(8):1026-31. |
| NHS Wales Questions to Assist with the Critical Appraisal of a Cross-Sectional Study | NHS Wales | Available from : https://www2.nphs.wales.nhs.uk/PubHObservatoryProjDocs.nsf/($All)/E7B0C80995DC1BA380257DB80037C699/$File/Cross%20sectional%20study%20checklist.docx?OpenElement. |
| Newcastle-Ottawa Scale | NOS | Available from : https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. |

**Appendix 3. Continued**

| Full name | Abbreviation | Reference |
|---|---|---|
| Academy of Nutrition and Dietetics ANDQ (Primary Research) | ANDQ | Available from : https://www.andeal.org/vault/2440/web/files/QCC_3.pdf. |
| Guidelines manual of National Institute for Health and Care Excellence: Appendices D-E | NICE | Available from : https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-appendices-bi-pdf-3304416006853. |
| Institute of Health Economic Quality Appraisal Tool for Case-Series Studies | IHE | Available from : https://www.ihe.ca/advanced-search/development-of-a-quality-appraisal-tool-for-case-series-studies-using-a-modified-delphi-technique. |
| Appraisal tool for Cross-Sectional Studies | AXIS | Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ open. 2016 Dec 1;6(12):e011458. |
| Agency for Healthcare Research and Quality Methodology Checklist | AHRQ | Viswanathan M, Patnode CD, Berkman ND, Bass EB, Chang S, Hartling L, Murad MH, Treadwell JR, Kane RL. Assessing the risk of bias in systematic reviews of health care interventions. Methods guide for effectiveness and comparative effectiveness reviews [Internet]. 2017 Dec 13. |
| NA | Pluye et al. 2009 | Pluye P, Gagnon MP, Griffiths F, Johnson-Lafleur J. A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in mixed studies reviews. International journal of nursing studies. 2009 Apr 1;46(4):529-46. |
| NA | Heller et al. 2008 | Heller RF, Verma A, Gemmell I, Harrison R, Hart J, Edwards R. Critical appraisal for public health: a new checklist. Public health. 2008 Jan 1;122(1):92-8. |
| CAse REport (CARE) Guidelines Checklist | CARE | CARE JJ, Kienle G, Altman DG, Moher D, Sox H, Riley D. The CARE guidelines: consensus-based clinical case reporting guideline development. Journal of medical case reports. 2013 Dec;7(1):1-6. |
| NA | Faillie et al. 2017 | Faillie JL, Ferrer P, Gouverneur A, Driot D, Berkemeyer S, Vidal X, Martínez-Zapata MJ, Huerta C, Castells X, Rottenkolber M, Schmiedl S. A new risk of bias checklist applicable to randomized trials, observational studies, and systematic reviews was developed and validated to be used for systematic reviews focusing on drug adverse events. Journal of Clinical Epidemiology. 2017 Jun 1;86:168-75. |
| Interventional Pain Management Techniques – Quality Appraisal of Reliability and Risk of Bias Assessment for Nonrandomized Studies | IPM-QRBNR | Manchikanti L, Hirsch JA, Heavner JE, Cohen SP, Benyamin RM, Sehgal N, Falco F, Vallejo R, Onyewu CO, Zhu J, Kaye AD. Development of an interventional pain management specific instrument for methodologic quality assessment of nonrandomized studies of interventional techniques. Pain physician. 2014;17(3):E291. |

**Appendix 3. Continued**

| Full name | Abbreviation | Reference |
|---|---|---|
| NA | Handu et al. 2016 | Handu D, Moloney L, Wolfram T, Ziegler P, Acosta A, Steiber A. Academy of Nutrition and Dietetics methodology for conducting systematic reviews for the Evidence Analysis Library. Journal of the Academy of Nutrition and Dietetics. 2016 Feb;116(2):311-8. |
| NA | Viswanathan et al. 2018 | Viswanathan M, Patnode CD, Berkman ND, Bass EB, Chang S, Hartling L, Murad MH, Treadwell JR, Kane RL. Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. Journal of clinical epidemiology. 2018 May 1;97:26-34. |
| NA | Young et al. 2009 | Young JM, Solomon MJ. How to critically appraise an article. Nature Clinical Practice Gastroenterology & Hepatology. 2009 Feb;6(2):82-91. |
| International Society of Pharmacoepidemiology Guidelines for Good Pharmacoepidemiology Practice | ISPE | Public Policy Committee, International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practice (GPP). pharmacoepidemiology and drug safety. 2016 Jan;25(1):2-10. |
| European Network of Centres for Pharmacoepidemiology and Pharmacovigilance Guide on Methodological Standards in Pharmacoepidemiology | ENCePP | Available from: https://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml. |
| Joint Task Force between the International Society for Pharmacoepidemiology and the International Society for Pharmacoeconomics and Outcomes Research | ISPE-ISPOR | Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, Gagne JJ, Gini R, Klungel O, Mullins CD, Nguyen MD. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1. 0. Value in health. 2017 Sep 1;20(8):1009-22. |
| Basque Office for Health Technology Assessment Tool | OSTEBA | Available from : http://www.lecturacritica.com/en/plataforma-flc_para-que-sirve-la-plataforma-web.php. |
| NA | Kennedy et al. 2019 | Kennedy CE, Fonner VA, Armstrong KA, Denison JA, Yeh PT, O'Reilly KR, Sweat MD. The Evidence Project risk of bias tool: assessing study rigor for both randomized and non-randomized intervention studies. Systematic reviews. 2019 Dec;8(1):1-0. |
| REporting of studies Conducted using Observational Routinely-collected Data Checklist | RECORD | Available from: https://www.record-statement.org/checklist.php. |

**Appendix 4. Ranking of appraisal tools based on the number of quality items addressed or sufficiently described**

| Ranking | Sufficiently described? | | Addressed by the tool? | |
|---|---|---|---|---|
| | **Appraisal tool** | **Number of items(%)** | **Appraisal tool** | **Number of items(%)** |
| ***Methodology*** | | | | |
| 1 | RTI Item Bank | 18 (69) | RTI Item Bank | 20 (77) |
| 2 | Faillie et al. 2017 | 8 (31) | Faillie et al. 2017 | 18 (69) |
| 3 | MINORS | 8 (31) | ANDQ | 13 (50) |
| 4 | ANDQ | 8 (31) | NICE | 12 (46) |
| 5 | ROBINS-I | 8 (31) | CER-CI | 11 (42) |
| 6 | Genaidy et al. 2007 | 6 (23) | CASP | 11 (42) |
| 7 | NIH | 6 (23) | SIGN | 11 (42) |
| 8 | CER-CI | 5 (19) | MINORS | 10 (38) |
| 9 | Handu et al. 2016 | 5 (19) | ROBINS-I | 10 (38) |
| 10 | NICE | 4 (15) | JBI | 10 (38) |
| 11 | JBI | 4 (15) | Blagojevic et al. 2010 | 10 (38) |
| 12 | GRACE | 4 (15) | Handu et al. 2016 | 9 (35) |
| 13 | Kennedy et al. 2019 | 4 (15) | AXIS | 9 (35) |
| 14 | ACROBAT-NRSI | 4 (15) | Heller et al. 2008 | 9 (35) |
| 15 | IPM-QRBNR | 3 (12) | IHE | 9 (35) |
| ***Reporting*** | | | | |
| 1 | STROBE | 13 (50) | STROBE | 17 (65) |
| 2 | TREND | 9 (35) | ENCePP | 15 (58) |
| 3 | Genaidy et al. 2007 | 8 (31) | TREND | 14 (54) |
| 4 | ENCePP | 8 (31) | RELEVANT | 12 (46) |
| 5 | ISPE | 6 (23) | Genaidy et al. 2007 | 10 (38) |
| 6 | Tseng et al. 2009 | 4 (15) | ISPE | 10 (38) |
| 7 | SURE | 3 (12) | SURE | 9 (35) |
| 8 | ISPE-ISPOR | 3 (12) | Tseng et al. 2009 | 7 (27) |

**Appendix 4. Continued**

| Ranking | Sufficiently described? | | Addressed by the tool? | |
|---|---|---|---|---|
| | Appraisal tool | Number of items(%) | Appraisal tool | Number of items(%) |
| ***Reporting*** | | | | |
| 9 | ANDQ | 3 (12) | ISPE-ISPOR | 7 (27) |
| 10 | RTI Item Bank | 3 (12) | ANDQ | 6 (23) |
| 11 | MINORS | 3 (12) | IHE | 6 (23) |
| 12 | RELEVANT | 2 (8) | Heller et al. 2008 | 6 (23) |
| 13 | IHE | 2 (8) | CER-CI | 5 (19) |
| 14 | CER-CI | 2 (8) | Rangel et al. 2003 | 5 (19) |
| 15 | Rangel et al. 2003 | 2 (8) | NICE | 5 (19) |

**Appendix 5. Ranking of appraisal tools based on the number of quality items on methodology, which were addressed or sufficiently described, segmented by quality domains**

| Ranking | Appraisal tool | Number of items(%) | |
|---|---|---|---|
| | | Sufficiently described? | Addressed? |
| ***Domain 1_Study design*** | | | |
| 1 | RTI Item Bank | 6 | 1 |
| 2 | ROBINS-I | 5 | 0 |
| 3 | NIH | 3 | 0 |
| 4 | CASP | 2 | 3 |
| 5 | IPM-QRBNR | 2 | 3 |
| 6 | ANDQ | 2 | 2 |
| 7 | ACROBAT-NRSI | 2 | 1 |
| 8 | Young et al. 2009 | 2 | 1 |
| 9 | Genaidy et al. 2007 | 2 | 1 |
| 10 | NHS Wales | 2 | 1 |
| 11 | Kennedy et al. 2019 | 2 | 0 |
| 12 | MINORS | 2 | 0 |
| 13 | Heller et al. 2008 | 1 | 4 |
| 14 | NICE | 1 | 3 |
| 15 | JBI | 1 | 2 |
| 16 | Faillie et al. 2017 | 1 | 2 |

**Appendix 5. Continued**

| Ranking | Appraisal tool | Number of items(%) | |
|---|---|---|---|
| | | Sufficiently described? | Addressed? |
| ***Domain 1_Study design*** | | | |
| 17 | AXIS | 1 | 2 |
| 18 | SURE | 1 | 1 |
| 19 | ISPE | 1 | 0 |
| 20 | GRACE | 1 | 0 |
| 21 | AHRQ | 1 | 0 |
| 22 | NOS | 1 | 0 |
| 23 | Weightman et al. 2004 | 1 | 0 |
| ***Domain 2_Data quality*** | | | |
| 1 | RTI Item Bank | 7 | 0 |
| 2 | ANDQ | 4 | 2 |
| 3 | MINORS | 4 | 2 |
| 4 | CER-CI | 3 | 1 |
| 5 | NIH | 3 | 1 |
| 6 | Faillie et al. 2017 | 2 | 5 |
| 7 | JBI | 2 | 2 |
| 8 | SIGN | 2 | 2 |
| 9 | Handu et al. 2016 | 2 | 1 |
| 10 | Kennedy et al. 2019 | 2 | 0 |
| 11 | GRACE | 2 | 0 |
| 12 | AHRQ | 2 | 0 |
| 13 | Genaidy et al. 2007 | 2 | 0 |
| 14 | Blagojevic et al. 2010 | 1 | 3 |
| 15 | ROBINS-I | 1 | 2 |
| 16 | GRADE | 1 | 2 |
| 17 | NOS | 1 | 2 |
| 18 | Thomas et al. 2004 | 1 | 2 |
| 19 | IPM-QRBNR | 1 | 1 |
| 20 | ACROBAT-NRSI | 1 | 0 |
| ***Domain 3_Data analysis*** | | | |
| 1 | Faillie et al. 2017 | 3 | 0 |
| 2 | MMAT | 2 | 1 |
| 3 | RTI Item Bank | 2 | 1 |

**Appendix 5. Continued**

| Ranking | Appraisal tool | Number of items(%) | |
|---|---|---|---|
| | | Sufficiently described? | Addressed? |
| ***Domain 3_Data analysis*** | | | |
| 4 | Viswanathan et al. 2018 | 2 | 0 |
| 5 | ROBINS-I | 2 | 0 |
| 6 | Weightman et al. 2004 | 1 | 2 |
| 7 | GRACE | 1 | 1 |
| 8 | Atluri et al. 2008 | 1 | 1 |
| 9 | CER-CI | 1 | 0 |
| 10 | NICE | 1 | 0 |
| 11 | Young et al. 2009 | 1 | 0 |
| 12 | MINORS | 1 | 0 |
| ***Domain 4_Results presentation*** | | | |
| 1 | Faillie et al. 2017 | 2 | 0 |
| 2 | Handu et al. 2016 | 2 | 0 |
| 3 | CER-CI | 2 | 0 |
| 4 | GRADE | 2 | 0 |
| 5 | ANDQ | 2 | 0 |
| 6 | RTI Item Bank | 1 | 0 |

**Appendix 6. Ranking of appraisal tools based on the number of quality items on reporting, which were addressed or sufficiently described, segmented by quality domains**

| Ranking | Appraisal tool | Number of items(%) | |
|---|---|---|---|
| | | Sufficiently described? | Addressed? |
| ***Domain 1_Study design*** | | | |
| 1 | STROBE | 5 | 2 |
| 2 | ENCePP | 4 | 4 |
| 3 | Genaidy et al. 2007 | 4 | 2 |
| 4 | RECORD | 4 | 1 |
| 5 | ISPE | 3 | 3 |
| 6 | Tseng et al. 2009 | 3 | 3 |
| 7 | TREND | 3 | 2 |
| 8 | Rangel et al. 2003 | 2 | 3 |
| 9 | ISPE-ISPOR | 2 | 1 |

**Appendix 6. Continued**

| Ranking | Appraisal tool | Number of items(%) | |
|---|---|---|---|
| | | Sufficiently described? | Addressed? |
| ***Domain 1_Study design*** | | | |
| 10 | Harm | 2 | 0 |
| 11 | MINORS | 2 | 0 |
| 12 | SURE | 1 | 3 |
| 13 | ANDQ | 1 | 3 |
| 14 | NIH | 1 | 2 |
| 15 | Montreal | 1 | 2 |
| 16 | CASP | 1 | 1 |
| 17 | CER-CI | 1 | 1 |
| 18 | IHE | 1 | 1 |
| 19 | AHRQ | 1 | 1 |
| 20 | NOS | 1 | 1 |
| 21 | Faillie et al. 2017 | 1 | 0 |
| 22 | RTI Item Bank | 1 | 0 |
| 23 | NHS Wales | 1 | 0 |
| ***Domain 2_Data quality*** | | | |
| 1 | STROBE | 5 | 0 |
| 2 | RECORD | 4 | 3 |
| 3 | TREND | 4 | 1 |
| 4 | ENCePP | 3 | 1 |
| 5 | ISPE | 3 | 0 |
| 6 | ANDQ | 2 | 0 |
| 7 | Genaidy et al. 2007 | 2 | 0 |
| 8 | ISPE-ISPOR | 1 | 3 |
| 9 | IPM-QRBNR | 1 | 0 |
| 10 | CARE | 1 | 0 |
| ***Domain 3_Data analysis*** | | | |
| 1 | STROBE | 2 | 2 |
| 2 | RELEVANT | 1 | 1 |
| 3 | CASP | 1 | 0 |
| 4 | JBI | 1 | 0 |
| 5 | SURE | 1 | 0 |

**Appendix 6. Continued**

| Ranking | Appraisal tool | Number of items(%) | |
|---|---|---|---|
| | | Sufficiently described? | Addressed? |
| ***Domain 3_Data analysis*** | | | |
| 6 | AXIS | 1 | 0 |
| 7 | CER-CI | 1 | 0 |
| 8 | Tseng et al. 2009 | 1 | 0 |
| 9 | Genaidy et al. 2007 | 1 | 0 |
| 10 | TREND | 1 | 0 |
| 11 | NHS Wales | 1 | 0 |
| ***Domain 4_Results presentation*** | | | |
| 1 | STROBE | 2 | 0 |
| 2 | IPM-QRBNR | 2 | 0 |
| 3 | IHE | 2 | 0 |

5

Chapter 6

# A Literature Review of Quality Assessment and Applicability to HTA of Risk Prediction Models of Coronary Heart Disease in Patients with Diabetes

Li Jiu, Junfeng Wang, Francisco Javier Somolinos-Simón, Jose Tapia-Galisteo, Gema García-Sáez, Mariaelena Hernando, Xinyu Li, Rick A Vreman, Aukje K Mantel-Teeuwisse, Wim G Goettsch

*Diabetes Research and Clinical Practice. 2024:111574*

## Abstract

This literature review had two objectives: to identify models for predicting the risk of coronary heart diseases in patients with diabetes (DM); and to assess model quality in terms of risk of bias (RoB) and applicability for the purpose of health technology assessment (HTA). We undertook a targeted review of journal articles published in English, Dutch, Chinese, or Spanish in 5 databases from 1st January, 2016 to 18th December, 2022, and searched three systematic reviews for the models published after 2012. We used PROBAST (Prediction model Risk Of Bias Assessment Tool) to assess RoB, and used findings from Betts et al. 2019 , which summarized recommendations and criticisms of HTA agencies on cardiovascular risk prediction models, to assess model applicability for the purpose of HTA. Of the 26 model studies and 30 models identified, only one model study showed low RoB in all domains, and no model was fully applicable for HTA. We advised that, to develop future models, the needs from HTA stakeholders, especially regarding health economics modelling, and the existing quality appraisal tools should be taken into account. Moreover, since general model applicability is not informative for HTA, novel adapted tools may need to be developed.

## Introduction

Coronary heart disease (CHD) is a heart disease featured by narrowing or blockage of coronary arteries (1,2). CHD is one of the major complications of diabetes mellitus, and it is diagnosed in more than one-fifth of patients with type-2 diabetes across all socioeconomic statuses (3). Previous reviews show that CHD is the leading cause of diabetes mortality, and it doubles the economic burden of patients with diabetes (DM) (4,5).

To reduce CHD morbidity, mortality, and relevant costs in DM patients, early recognition of high CHD risks and appropriate selection of prevention strategies are highly needed (6,7). To achieve this, risk prediction models have been widely used by clinicians to estimate probabilities of the occurrence of CHD in DM patients (8,9). Apart from their clinical use, risk prediction models can be applied for the purpose of health technology assessment (HTA). For example, by functioning as a part of a cost-effectiveness model, i.e., allowing detailed exploration of heterogeneous outcomes among different subpopulations of interest to decision makers, risk prediction models can be utilized by HTA stakeholders to directly estimate clinical and economic impact of a health intervention (10-15).



As risk prediction models for DM patients started to emerge, the variety in techniques used for developing the models increased as well. The most frequently used technique is statistical modelling, which can be further categorized into Cox regression analysis, Logistic regression analysis, Weibull regression analysis, Gompertz regression analysis, etc. (16,17). Another modelling technique that emerges is machine learning (ML), including but not limited to neural networks, random forest, decision-tree, support vector machine, etc. (17). ML models are gaining popularity in the field of diabetes due to their capability to capture the complex relationships among a vast number of predictors (18), which provides a potential for better predictive performance.

However, the ever-increasing number of models and variety of modelling techniques have placed a heavy burden on model evaluation and raised concerns about model quality in terms of risk of bias (RoB) and applicability. A high RoB, which is pervasive among risk prediction models (19), could increase the likelihood that models yield inaccurate prediction, and decrease the confidence of users (e.g. clinicians, HTA researchers and agencies, patients, etc.) in model performance (20). In addition, model users are at risk of selecting and applying a suboptimal model for their own purposes, as they often miss information on therapeutic, geographic, or temporal settings in which the models can be applied (21).

To assess quality of CHD risk prediction models for DM patients, Van Dieren et al. conducted a systematic review to summarize the structure and predictive performance of existing models for predicting type 2 diabetes published before May 2012 (16). More recently, Galbete et al. updated the review conducted by Van Dieren et al. by searching for the models published before July 2021 (22). They summarized model performance and assessed RoB and generic model applicability. However, these two reviews adopted similar search strategies using a single data source (either Medline or PubMed) and did not systematically search for risk prediction models developed with ML techniques. Also, they did not provide an assessment of model applicability for the purpose of HTA, which requires special considerations, such as appropriateness of subgroup populations (14).

Hence, the aim of our study was to identify the risk prediction models developed recently with statistical or machine learning techniques, and to assess their RoB and applicability for HTA. This research was performed as part of the HTx project (22). The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825162.

## Methods

### Protocol

A research protocol of this study was registered in the PROSPERO database (ID CRD42021273240), then rigorously followed. To conduct the systematic review, we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (23).

### Data source & Search strategy

We searched PubMed, Embase, Scopus, Web of Science, and IEEE Xplore for journal articles and conference papers predicting CHD risks in DM patients, in two rounds (published from 1st January, 2016 to 31st May, 2021; published from 1st June, 2021 to 18th December, 2022). We used a search strategy (Appendix 1) with three concepts (i.e. risk prediction, CHD, and diabetes), which was developed from published strategies to retrieve relevant publications for CHD (24,25) or prediction models (26,27). The search strategy was developed by two reviewers (LJ & JW), then edited by an experienced librarian in document retrieval from Utrecht University. We also checked citations in all identified relevant studies. In addition to the database search, we searched three recently published systematic reviews (i.e. Galbete et al. 2022 (21), Faizal et al. 2021 (28),

and Lenselink et al. 2022 (29) ) which identified models predicting the risk of cardiovascular diseases, in DM patients or general population.

### Eligibility criteria

A study was eligible if (1) it described the development of a prediction model; (2) the target population was patients with diabetes; (3) the outcome of prediction was CHD or a CHD component (i.e. myocardial infarction, acute coronary syndrome, and/or angina); (4) the study was published in English, Dutch, Spanish, or Chinese, based on the review team's language proficiencies; (5) the study was published after 2012; (6) the full-text was available. Exclusion criteria included non-human studies or studies only describing model validation. Studies using heart or cardiovascular disease as a combined outcome only were also excluded because the risk of CHD could not be predicted.

### Study selection & Data collection

For study selection, one reviewer (LJ) screened titles and abstracts of all identified studies, while another (GGS) independently screened a random set of 20%. Then two reviewers (LJ and FJSS or JTG) independently scanned full texts of studies that might be eligible. Any disagreement between reviewers was solved through consensus. For each model identified, one reviewer (LJ) extracted model characteristics (e.g. target population, outcome, and number of predictors, etc.), based on a data collection form developed from previous reviews (15,21).



### Quality assessment

For assessing RoB, several appraisal tools can be used, such as the PROBAST (Prediction model Risk Of Bias Assessment Tool) (19), CHARMS (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) (30), TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) (31), and STROBE (Strengthening the Reporting of Observational studies in Epidemiology) (32). We determined to use the PROBAST, because it was carefully and specifically designed for evaluating RoB of prediction models. It consists of four RoB domains (i.e. participants, predictors, outcomes, and analysis), and each domain-specific RoB is graded as low, high, or unclear.

For assessing model applicability for the HTA purpose, we did not use PROBAST, because it mainly addressed applicability concerns regarding medical settings, i.e., whether population, predictors, or outcomes of a study differed from those specified in a systematic review question (20). Additionally, we did not find any specifically designed tool but only a review conducted by Betts et al. (15), which

summarized reasons why HTA agencies recommended or criticized models for predicting cardiovascular diseases. Betts et al. mentioned three aspects of concern regarding applicability for the purpose of HTA, including geographic and therapeutic generalizability, whether the model was up-to-date, and appropriateness of model covariates. According to Betts et al., seven signaling questions (SQs) were formulated by two reviewers (LJ and JW), and then edited by five reviewers (GGS, FJSS, JTG, XL, LJ) after a pilot quality assessment of one-third of the eligible models. The questions and their rationales are attached in Appendix 2.

Quality assessment was conducted independently by two reviewers (LJ and FJSS, JTG, or XL), and any discrepancy was solved through discussion with at least three reviewers. Before the formal RoB assessment, two training sessions with three example modelling studies (e.g. Covid-19) were conducted by six reviewers (JW, LJ, FJSS, JTG, GGS).

### Data analysis

For data analysis, we narratively synthesized characteristics of the eligible models by presenting all variables as numbers and percentages. The results were presented separately for ML and statistical models in both tables and graphs.

## Results

### Model selection

Among a total of 12784 records identified from the five databases, 1381 were eliminated as duplicates, leaving a total of 11403 initial records (Figure 1). After adding records from the three published reviews and reviewing titles and abstracts, we selected 183 records for full-text screening, then excluded 157 records with the reasons such as inappropriate population (n=58). No new references were obtained through the reference lists of the remaining articles. Therefore, 26 studies, which described 30 models, were included for data extraction. Twenty-one, three, and two of the studies were identified from database search only, published reviews only, or both. Since RoB of the three studies (33-35), which were identified from the published reviews, had been previously assessed (21,36), we included 23 model studies for RoB assessment and all the 30 models for applicability assessment.

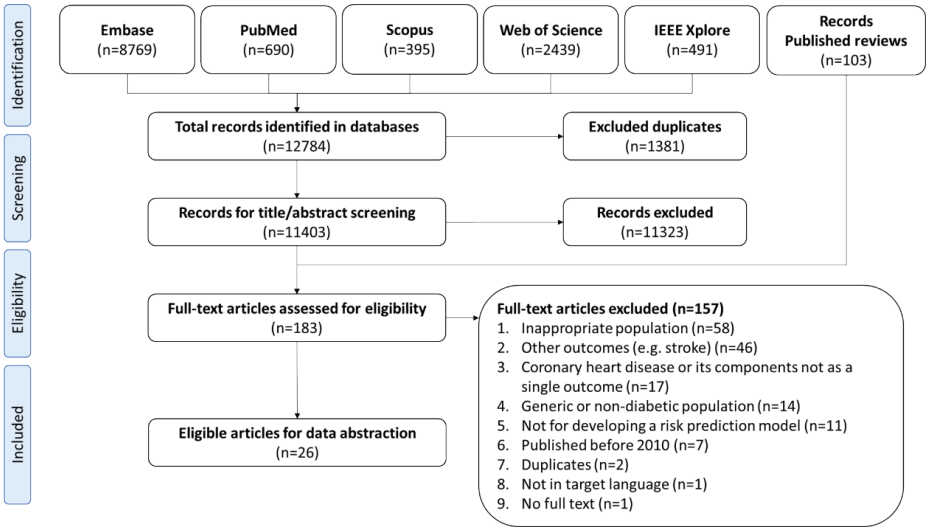



**Figure 1. Flowchart of included studies.**

## Model Characteristics

The summary of study and model characteristics is presented in Table 1, and the reference list and model details can be found in Appendix 3 & 4.

**Table 1. Characteristics of eligible models**

| | Number | Percentage (%) |
|---|---|---|
| **Study characteristics (n=26)** | | |
| ***Study design*** | | |
| Observational study[a] | 21 | 81 |
| Trial + Observational study | 3 | 12 |
| Trial[b] | 2 | 7 |
| ***Disease*** | | |
| Type-2 diabetes | 22 | 85 |
| Type-1 diabetes | 1 | 4 |
| All diabetes[c] | 3 | 11 |

**Table 1. Continued**

| | Number | Percentage (%) |
|---|---|---|
| **Study characteristics (n=26)** | | |
| ***Sample size of model development*** | | |
| >10000 | 6 | 23 |
| 1000-10000 | 13 | 50 |
| <1000 | 7 | 27 |
| ***Region*** | | |
| Asia | 12 | 46 |
| North America | 8 | 31 |
| Europe | 5 | 19 |
| Intercontinental | 1 | 4 |
| ***Method of internal validation*** | | |
| Cross-validation | 5 | 19 |
| Bootstrapping | 5 | 19 |
| Sample split | 8 | 31 |
| Sample split $ Cross-validation | 1 | 4 |
| Not reported | 7 | 27 |
| ***External validation conducted?*** | | |
| Yes | 10 | 38 |
| No | 16 | 62 |
| **Model characteristics (n=30)** | | |
| ***Model type*** | | |
| Machine learning[d] | 5 | 17 |
| Cox | 17 | 57 |
| Logistic | 4 | 13 |
| Other (e.g. Linear) | 4 | 13 |
| ***Outcome of interest[e]*** | | |
| CHD | 16 | 53 |
| MI | 11 | 37 |
| ACS | 3 | 10 |
| ***Age of simulated individuals*** | | |
| All | 17 | 57 |
| With a range (e.g. 40-64) | 8 | 27 |
| Not reported | 5 | 16 |
| ***Time cycle of prediction*** | | |
| > 1 year | 12 | 40 |
| 1 year at maximum | 4 | 13 |
| Not reported | 14 | 47 |

**Table 1. Continued**

| | Number | Percentage (%) |
|---|---|---|
| **Model characteristics (n=30)** | | |
| ***Number of final predictors*** | | |
| <=10 | 12 | 40 |
| >10 | 11 | 37 |
| Not reported | 7 | 23 |

[a] "Observational study" indicates the data were derived from one or several previous observational studies, or from one or several databases, such as electronic health/medical records, registry, or administrative claims data.

[b] "Trial" indicates the data were derived from one or several previous trials.

[c] "All diabetes" indicates that the study did not specify the type of diabetes, such as Type 1, Type 2, gestational diabetes.

[d] "Machine learning" indicates the machine learning techniques used for developing the included prediction models, including Multi-task Learning (MTL), Random Forest (RF), Neural Network (NN) , and Recurrent Neural Network Gated Recurrent Unit (RNN GRU).

[e] "CHD" indicates coronary heart disease, "MI", myocardial infarction, "ACS", acute coronary syndrome.



Mostly, the data used for model development were derived from observational studies (21;81%). Regarding target population, most models focused on only patients with type-2 diabetes (22;85%), and sourced patients from a single continent, that is, Asia (12;46%), North America (eight;31%), and Europe (five;19%), while the one left sourced from four continents (37). For the sample size, most studies (19; 73%) had a number larger than 1000, six of which had sizes even larger than 10000.

Regarding model characteristics, models developed with the statistical approaches, such as Cox (17,57%) and Logistic (4, 13%), accounted for the most, while the other five used various ML techniques, that is, Multi-task Learning (38), Random Forest (39), Neural Network (40), Recurrent Neural Network Gated Recurrent Unit (41), and K-nearest Neighbor models (42). Regarding outcomes, about half of the models predicted CHD, while other predicted MI or acute coronary syndrome. In addition, almost 60% models could predict disease risk in patients of all ages, while 5 models did not report the age range. Also, only 16 (53%) models reported the duration of risk they could predict (e.g. 5-year CHD risk), and only 4 models provided the equations to predict the annual risk.

## RoB Assessment - PROBAST

The quality assessment in terms of RoB is shown in Figure 2. In Appendix 5 & 6, the results were splitted if the model studies were developed with statistical or ML techniques.

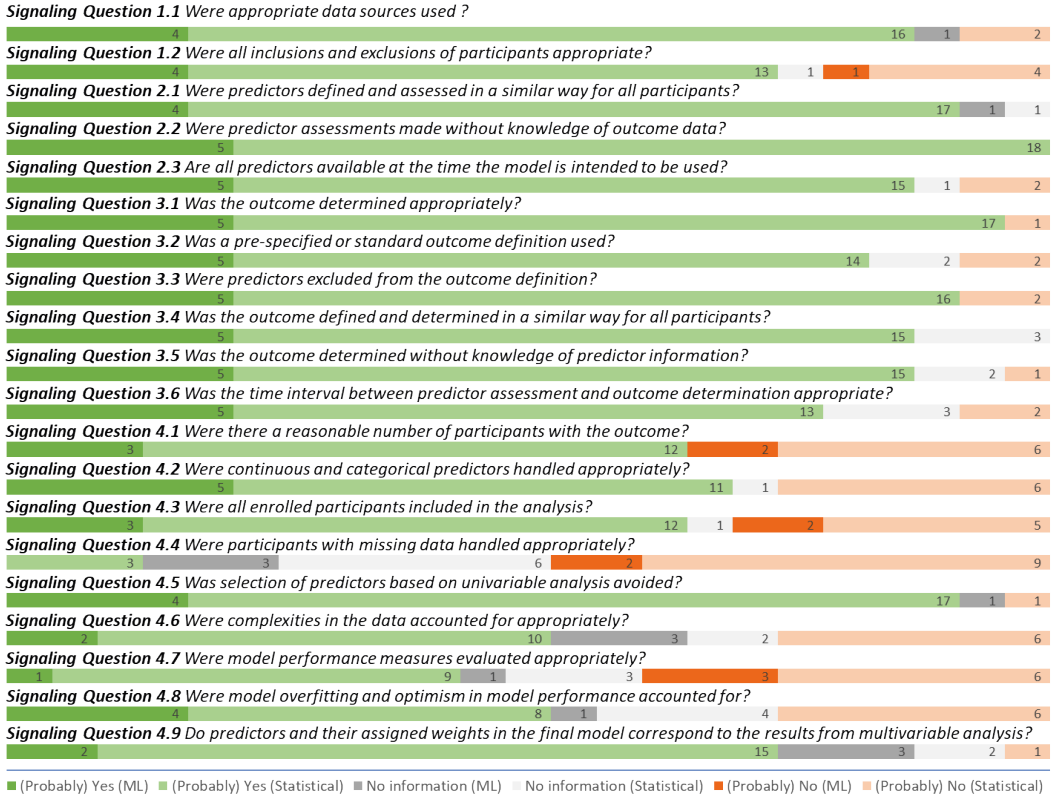



**Figure 2. PROBAST signaling questions for the 23 studies investigating models developed with statistical or machine learning techniques.**
ML indicates machine learning models; Statistical indicates statistical models. The signaling questions 1.1-1.2, 2.1-2.3, 3.1-3.6, and 4.1-4.9 corresponds to the risk of bias domain of participants, predictors, outcome, and analysis, respectively.

In the Participants domain (SQ 1.1-1.2), three (60%) ML model studies and 12 (67%) statistical model studies were rated as having low RoB. Appropriate data sources, such as prospective cohort and randomized controlled trial data, were used (SQ 1.1) in four (80%) ML model studies and 16 (89%) statistical model studies. The included patients were considered representative of the target population (SQ 1.2) in four (80%) ML model studies and 13 (72%) statistical model studies. A total of four models included patients who were already known to have the CHD-related outcomes at the time of predictor measurement (e.g. patients with CHD history), and one model excluded sicker patients based on number of hospitalization (41). Consequently, the predictive performance of these models could be overestimated or underestimated, respectively.

In the Predictors domain (SQ 2.1-2.3), four (80%) ML model studies and 15 (83%) statistical model studies were rated as low RoB. Predictors were defined and assessed in a similar way for all participants (SQ 2.1) in four (80%) ML model studies and 17 (94%) statistical model studies. Predictor assessments were made without knowledge of outcome data (SQ 2.2) in 21(91%) studies. All predictors were considered available at the time the model is intended to be used (SQ 2.3) in all ML models, but only in 15 (83%) statistical models. The remaining three statistical models were considered high RoB, because two of them included predictors that were unlikely to be available in clinical practice (e.g. anthropometric measurement) (37,43), and one did not mention when the model would be used (44).



In the Outcome domain (SQ 3.1-3.6), all the ML model studies (100%) were rated as low RoB, and were considered high-quality in Signaling questions from 3.1 to 3.6. Comparably, 10 (56%) statistical model studies were rated as low RoB. Only one statistical model (41) did not use appropriate methods to determine the outcome, thus increasing the risk of misclassification (SQ 3.1). Similarly, only two statistical model studies (45,46) missed prespecified or standard definitions to determine the outcome (SQ 3.2). Likewise, predictors were mistakenly included in the outcome definition (SQ 3.3) in two statistical model studies (47,48). Outcomes were defined and measured in a similar way (SQ 3.4) in 15 (83%) statistical model studies, except three which provided no information (49-51). According to SQ 3.5, prediction information was known only in one statistical model (47) when determining the outcome status. In SQ 3.6, the time interval between predictor assessment and outcome determination was considered too short in two statistical models (43,45).

In the Analysis domain (SQ 4.1-4.9), most concerns regarding RoB were identified. All the ML model studies and 16 (89%) statistical model studies

were rated as high RoB. SQ 4.7 showed that three (60%) ML model studies and six (33%) statistical model studies did not appropriately evaluate model performance, because they missed calibration evaluation (39,41,44,48,51-54), only used the Hosmer–Lemeshow test for calibration evaluation (55), or presented classification measures with predicted probability thresholds derived from the data set at hand (40). According to SQ 4.4, two (40%) ML model studies (38,42) and nine (50%) statistical model studies handled missing data inappropriately by simply excluding them, while three (60%) ML model studies (39-41) and six (33%) statistical model studies suffered from no information. SQ 4.2 revealed that continuous and categorical predictors were handled appropriately in all ML model studies, but only in 11 (61%) statistical models. Six (33%) studies did not examine non-linearity for continuous variables (47,52,54,56) or categorized continuous variables (52,55). Similarly, model overfitting and optimism were considered (SQ 4.8) in four (80%) ML model studies, but only in eight (44%) statistical model studies. Six (33%) studies did not use internal validation techniques (44,52-56), or the validation did not include the whole model development procedures (43,45). According to SQ 4.3, two (40%) ML models (38,41) and five (28%) statistical models (44,45,47,54,57) inappropriately excluded patients due to uninterpretable findings, outliers, or missing data. SQ 4.6 finally showed that none of the ML model studies but six statistical model studies inappropriately addressed censoring or competing risks. Three models simply used logistic regression for censoring (47,52,56) and three ignored competing risks (43,52,57). Also, five studies provided no information. In response to SQ 4.1, two ML model (39,42) and six (33%) statistical models did not have a reasonable number of participants with the outcome (i.e. event per predictor parameter < 10).

The remaining questions contributed relatively less to the overall RoB. Only one model (55) selected predictors based on univariable analysis, and another (40) provided no information (SQ 4.5). Based on SQ 4.9, information was missing on whether predictors and their assigned weights in the final model correspond to the results from multivariable analysis in three (60%) ML models (38,40,41) and two (11%) statistical models (47,56). Additionally, in only one statistical model (55), the predictors did not correspond to the results.

## Applicability to HTA Assessment

The assessment in terms of applicability for the purpose of HTA is shown in Figure 3 and Appendix 7.

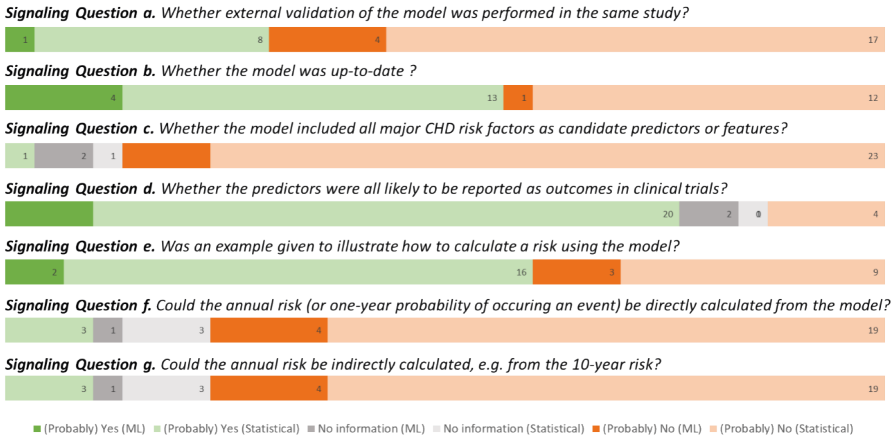



**Figure 3. Model applicability for the purpose of HTA.**
ML indicates machine learning models; Statistical indicates statistical models.

In summary, the applicability of the models for HTA was quite limited, as none of the 30 models were fully applicable (i.e. "Yes" or "Probably Yes" in all the seven signaling questions). Only six models in three studies (33,57,58) had an "Yes" or "Probably Yes" in at least four signaling questions. The major barrier of model applicability was the lack of feasibility to calculate the annual risk of CHD or its component, either directly or indirectly (SQ f & SQ g), as only three models (33,58) could provide the option. The direct calculation indicates that, an equation or tool (e.g. an online user interface) is provided in the original study to calculate the annual risk of disease. The indirect calculation indicates that, though equations or a tool for predicting the annual risk are not provided, users could calculate the risk, using evidence provided in the original study (e.g. using hazard functions to calculate the accumulated risk). Another barrier of model applicability was inappropriate exclusion of major CHD risk factors as candidate predictors (SQ d),as only one model (44) considered all the factors as candidate predictors. The CHD risk factors or features refer to those defined by a recently published overview, Hajar 2017 (59), including blood pressure, high blood cholesterol levels, smoking, overweight or obesity, lack of physical activity, unhealthy diet and stress, age, gender, family history, and race . Additionally, external validation was performed within the same study in only one ML (38) and eight statistical models (SQ a). Although all the identified models were published from 2013 onwards, one ML (32)

or 12 statistical models were considered relatively obsolete (SQ b), because all the follow-ups of their target populations ended before 2012. Additionally, about one-third of models were attached with examples on how these models could be used. For example, the model from the United Kingdom Prospective Diabetes Study (UKPDS) (33) and from Ye et al. 2022 (58) created an artificial patient with assumed value on its characteristics, and illustrated how the CHD-related risk was calculated using the model. Finally, according to SQ c, four (22%) statistical models (44,49,50,56) included predictors that were not likely to be reported as trial outcomes, such as biomarkers.

# Discussion

## Findings

We conducted a literature review of models which predicted CHD risk in patients with diabetes to assess quality, in terms of risk of bias, and applicability for the purpose of HTA. We identified 25 statistical and five ML models, with overall relatively poor model quality. Only one study (60) showed low RoB in all domains of the PROBAST checklist, and none of the 30 models were fully applicable for HTA.

We discovered that most of the major contributors of high RoB were located in the analysis domain. Similar findings were also reported by Galbete et al., Haider et al., and Van der Heijden et al., who assessed RoB of 65, 14, and 16 models, respectively, for predicting the risk of cardiovascular disease or retinopathy in general or DM populations (22,36,61). This finding is as expected, because the analysis domain, which addresses statistical considerations of model development, is the most complicated, with the most (n=9) signaling questions (19). We did not identify similar research that assessed the model applicability for the HTA purpose.

The overall high RoB of the identified model studies implied that PROBAST might not be strictly followed by model developers. To go one step further, we could speculate that, model developers did not fully comply with some other published appraisal tools either, because, as mentioned in the method part, these tools also highlighted similar RoB concerns that were not adequately addressed by the identified models. For example, the CHARMS discouraged the complete-case analysis for addressing missing data, emphasized the importance of recoding model performance in terms of calibration and discrimination, and recommended the use of bootstrapping and cross-validation against overfitting (30). Also, the STROBE guideline emphasized the

importance of reporting missing data in modelling studies (32). In contrast, almost half of the identified model studies reported no information regarding missing data.

One explanation for the lack of compliance might be the failure of disseminating the appraisal tools. The publications that described the successful external validation cannot prove the success of dissemination, as the potential model developers who do not understand an appraisal tool would never use it or describe their confusion in their own modelling study. Alternatively, the lack of compliance may be attributed by the lack of feasibility to apply the tools. For example, although all the above-mentioned tools discouraged the use of complete case analysis for addressing missing data, the complete case analysis might not lead to biases. In certain conditions, it achieved precision similar to or better than multiple imputation, and high statistical coverage (62). If this was true, the existing tools might need to be adapted to approve the use of complete case analysis in some scenarios. Hence, further research may be needed to analyze whether the model developers have understood the existing tools, and how they have used them. We expect that future research could contribute to improved appraisal tools and the relevant dissemination strategies.



Additionally, it seems that most developers of risk prediction models did not fully understand the needs for the HTA purpose, so the potential of these models was not fully explored. For example, health economics models in the diabetes field, especially for those with a Markov structure, often need empirical data or risk prediction models for predicting the CHD risk, with an aim to accurately calculate the overall costs and effectiveness of a cohort. Compared to aggregate data, risk prediction models could enable the estimation of cost-effectiveness at individual level, and they are a good alternative to empirical patient data from real-world databases (63,64). However, to apply risk prediction models to health economics modelling, the original mathematical equations should be provided and called repeatedly. Our results showed that, some studies only provided an online user interface without an equation, which could not satisfy the relevant needs (43,52,58). Also, it is worth noting that, Cox models are a popular type of risk prediction models for the HTA purpose, as they could predict time to an event and an event risk within a time interval of any length. In particular, Cox models are suitable to discrete-event simulation models, an increasingly popular health economics model featured by great flexibility to handle time-to-event data (64). However, most of our identified Cox models (n=17) were not applicable. The reason was that, they only provided a cumulative hazard function with fixed model coefficients, for estimating the CHD-related risk for 3,5,or 10 years, without providing the original hazard function, which enabled estimation of an instantaneous risk. These models could satisfy the needs for clinical decision-making, as information on a 5- or 10-year

event risk, could help health-care providers or patients decide on which treatment to receive. However, these models could not be incorporated into a health economics model, unless assumptions are made on the instantaneous event risk (e.g. constant overtime), which would increase ROB. Therefore, we highly recommend developers of risk prediction models not to develop, but to improve their existing models, by reporting their mathematical equations more transparently, or by at least providing a cumulative function that could predict an annual event risk.

Another finding regarding model applicability for HTA was that, although all the models included some CHD risk factors as model covariates, they were not in full agreement on which risk factors should be included. For example, while all the models included age, sex, and smoking as covariates, they differed in whether to include diet, physical activity, or mental health. The appropriate inclusion of CHD risk factors as model covariates has been considered by HTA agencies as evidence of wide model applicability. For example, the Dental and Pharmaceutical Benefits Agency in Sweden and the Dutch Healthcare Insurance Board in the Netherlands commented on the absence of any cholesterol measure as a covariate in a cardiovascular risk prediction model called REACH. (15) However, we identified no clear statement from HTA agencies, or even from clinical guidelines, on what risk factors should at least be included as model covariates. Indeed, many studies have investigated the issue by providing a list of major CHD risk factors (65-68), but their recommendations vary. Consequently, the lack of agreement on CHD risk factors to be included in a risk prediction model would confuse model developers, and ultimately reduce model applicability. Hence, we suggest developing a generic framework which summarizes clinical risk factors as model covariates in the diabetes field. The framework may not only address risk factors of CHD, but also those of other major DM complications.

We found that the concerns regarding model applicability for HTA cannot be simply addressed by the assessment of generic applicability. As mentioned by PROBAST, the generic applicability considers the extent to which the population, outcome, and definition and assessment of predictors match a review question. However, the generic applicability doesn't imply much regarding how to develop a model with wide applicability, as the PROBAST could not expect what review questions can be imposed by the HTA stakeholders. Consequently, the model users might only select and apply the least unsatisfactory model, while losing the opportunity of acquiring a perfect one. One solution for this applicability concern is to account for needs of HTA stakeholders in appraisal tools. This could be achieved by adapting existing appraisal tools or developing new tools. However, given the various needs of model users, innovating an one-size-fits-all appraisal tool which defines an one-size-fits-all risk prediction

model may not be feasible. Therefore, to account for various needs, we recommend closer collaboration among model developers, tool developers, and HTA stakeholders, and suggest the involvement of all stakeholders in development and implementation of appraisal tools.

### Comparison of ML and statistical models

Since the numbers of ML and statistical models we identified are small, we could not compare quality of the two. However, we, as reviewers, feel that it is harder to assess quality of ML models than statistical models. One obvious reason is that ML models include unique features that could not be highlighted by generic quality appraisal tools. For example, ML models might have built-in capabilities for handling missing data (51). To address this concern, several tools specifically designed for ML models are being developed, such as the PROBAST-AI and STROBE-AI (69). Another reason for the difficult quality assessment is that ML models normally adopt a black box approach that prevent users from interpreting the reasoning behind a models' prediction (70). To address this concern, a research topic – Explainable AI – has emerged, and novel approaches for improved interpretability are being developed (71). However, as model users often need to compare quality of models developed with various techniques, we suggest exploring methods to compare quality of statistical and ML models while taking into account their particularities.



### Limitations

Our study still has limitations. One limitation is that we might have missed models, as only one reviewer scanned all titles and abstracts, while another scanned a random set of 20%. However, tracking references of included studies did not yield additional references. Our findings regarding overall model quality were supported by other studies and will not be disturbed by the potentially missing models. Another limitation is that our results regarding model applicability for the purpose of HTA is explorative, and the evaluation criteria were from a single review (i.e. Betts et al (15)) and authors' opinions. While our results cover key concerns of HTA stakeholders, some concerns may not be covered. Hence, extra efforts are needed if HTA stakeholders apply the models based on our results.

## Conclusions

Both models based on machine learning and statistical techniques have been developed to predict the CHD risk in DM patients, but the quality, in terms of risk of bias and model applicability for the purpose of HTA, is relatively low. Model developers mostly did not understood the needs from HTA stakeholders, and we recommend further research to explore the reasons. In addition, novel tools are needed, as the existing tools which only address generic model applicability could not satisfy the needs of HTA stakeholders. To achieve this, model developers, tool developers, and HTA stakeholders may need closer collaboration.

## Author contribution

LJ designed the search strategy, scanned all hits, conducted full-text review of potential eligible studies, collected data, participated in quality assessment of risk prediction models, analyzed and interpreted data, and wrote the manuscript. All co-authors contributed to this study and critically reviewed and approved the manuscript.

# References


1. National Cancer Institute: Coronary heart disease. Available from : https://www.cancer.gov/publications/dictionaries/cancer-terms/def/coronary-heart-disease. [Accessed Jun 11, 2022].
2. Centers for Disease Control and Prevention: Coronary Artery Disease (CAD). Available from: https://www.cdc.gov/heartdisease/coronary_ad.htm. [Accessed Jun 11, 2022].
3. Einarson TR, Acs A, Ludwig C, Panton UH. Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017. Cardiovasc Diabetol. 2018 Dec;17(1):1-9.
4. Khalil CA, Roussel R, Mohammedi K, Danchin N, Marre M. Cause-specific mortality in diabetes: recent changes in trend mortality. Eur J Prev Cardiol. 2012 Jun 1;19(3):374-81.
5. Einarson TR, Acs A, Ludwig C, Panton UH. Economic burden of cardiovascular disease in type 2 diabetes: a systematic review. Value Health. 2018 Jul 1;21(7):881-90.
6. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol. 2019 Sep 10;74(10):e177-232.
7. Marshall T. Coronary heart disease prevention: insights from modelling incremental cost effectiveness. BMJ. 2003 Nov 27;327(7426):1264.
8. Van Der Heijden AA, Ortegon MM, Niessen LW, Nijpels G, Dekker JM. Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. Diabetes Care. 2009; 32(11): 2094-2098.
9. Bhopal R, Fischbacher C, Vartiainen E, Unwin N, White M, Alberti G. Predicted and observed cardiovascular disease in South Asians: application of FINRISK, Framingham and SCORE models to Newcastle Heart Project data. J Public Health (Oxf) 2005; 27(1): 93-100.
10. Stevanovic J, Postma MJ, Pechlivanoglou P. A systematic review on the application of cardiovascular risk prediction models in pharmacoeconomics, with a focus on primary prevention. Eur J Prev Cardiol. 2012; 19(2_suppl): 42-53.
11. Palmer AJ, Roze S, Valentine WJ. The CORE Diabetes Model: projecting long-term clinical outcomes, costs and cost-effectiveness of interventions in diabetes mellitus (types 1 and 2) to support clinical and reimbursement decision-making. Curr Med Res Opin. 2004 Jan 1;20(sup1):S5-26.
12. Eddy DM, Schlessinger L. Archimedes: a trial-validated model of diabetes. Diabetes Care. 2003; 26(11): 3093-3101.
13. Mueller E, Maxion-Bergemann S, Gultyaev D, et al. Development and validation of the Economic Assessment of Glycemic Control and Long-Term Effects of diabetes (EAGLE) model. Diabetes Technol Ther. 2006; 8(2): 219-236.
14. Betts MB, Milev S, Hoog M, et al. Comparison of recommendations and use of cardiovascular risk equations by health technology assessment agencies and clinical guidelines. Value Health. 2019 Feb 1;22(2):210-9.
15. Van Dieren S, Beulens JW, Kengne AP, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. Heart 2012; 98(5): 360-369.
16. Xu Q, Wang L, Sansgiry SS. A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning. J Med Artif Intell. 2020; 3(6).

17. Kwon O, Na W, Kim YH. Machine learning: a new opportunity for risk prediction. Korean Circ J. 2020 Jan 1;50(1):85-7.

18. Venema E, Wessler BS, Paulus JK. Large-scale validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. J Clin Epidemiol. 2021;138: 32-39.

19. Moons KG, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. 2019 Jan 1;170(1):W1-33.

20. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. J Clin Epidemiol. 2016 Nov 1;79:76-85.

21. Galbete A, Tamayo I, Librero J, Enguita-Germán M, Cambra K, Ibáñez-Beroiz B. Cardiovascular risk in patients with type 2 diabetes: A systematic review of prediction models. Diabetes Res Clin Pract. 2021: 109089.

22. HTx: About HTx project. Available from: https://www.htx-h2020.eu/about-htx-project. [Accessed 2022 Oct 25].

23. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

24. De Waure C, Lauret GJ, Ricciardi W, et al. Lifestyle interventions in patients with coronary heart disease: a systematic review. Am J Prev Med. 2013 Aug 1;45(2):207-16.

25. Wolters FJ, Segufa RA, Darweesh SK, et al. Coronary heart disease, heart failure, and the risk of dementia: a systematic review and meta-analysis. Alzheimers Dement. 2018 Nov 1;14(11):1493-504.

26. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc. 2001 Jul 1;8(4):391-7.

27. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PloS One. 2012 Feb 29;7(2):e32844.

28. Faizal AS, Thevarajah TM, Khor SM, Chang SW. A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach. Comput Methods Programs Biomed. 2021 Aug 1;207:106190.

29. Lenselink C, Ties D, Pleijhuis R, van der Harst P. Validation and comparison of 28 risk prediction models for coronary artery disease. Eur J Prev Cardiol. 2022 Mar 1;29(4):666-74.

30. Moons KG, de Groot JA, Bouwmeester W et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PloS Med. 2014; 11(10): e1001744.

31. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Surg. 2015;102(3): 148-158.

32. Vandenbroucke JP, Von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PloS Med. 2007;4(10): e297.

33. Hayes AJ, Leal J, Gray AM, et al. UKPDS outcomes model 2: a new version of a model to simulate lifetime health outcomes of patients with type 2 diabetes mellitus using data from the 30 year United Kingdom Prospective Diabetes Study: UKPDS 82. Diabetologia. 2013 Sep;56:1925-33.

34. Nishimura K, Okamura T, Watanabe M, et al. Predicting coronary heart disease using risk factor categories for a Japanese urban population, and comparison with the randomized risk score: the suita study. J Atheroscler Thromb. 2014 Aug 26;21(8):784-98.

35. Piniés JA, González-Carril F, Arteagoitia JM, et al. Development of a prediction model for fatal and non-fatal coronary heart disease and cardiovascular disease in patients with newly diagnosed type 2 diabetes mellitus: the Basque Country Prospective Complications and Mortality Study risk engine (BASCORE). Diabetologia. 2014 Nov;57:2324-33.

36. Haider S, Sadiq SN, Moore D, Price MJ, Nirantharakumar K. Prognostic prediction models for diabetic retinopathy progression: a systematic review. Eye. 2019 May;33(5):702-13.

37. Rådholm K, Chalmers J, Ohkuma T, et al. Use of the waist-to-height ratio to predict cardiovascular risk in patients with diabetes: R esults from the ADVANCE-ON study. Diabetes Obes Metab. 2018 Aug;20(8):1903-10.

38. Kim E, Caraballo PJ, Castro MR, Pieczkiewicz DS, Simon GJ. Towards more accessible precision medicine: building a more transferable machine learning model to support prognostic decisions for micro-and macrovascular complications of type 2 diabetes mellitus. J Med Syst. 2019 Jul;43(7):1-2.

39. Fan R, Zhang N, Yang L, Ke J, Zhao D, Cui Q. AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus. Sci Rep 2020. Sep 2;10(1):1-8.

40. Longato E, Fadini GP, Sparacino G, Gubian L, Di Camillo B. Prediction of cardiovascular complications in diabetes from pharmacy administrative claims. In2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON) 2020 Jun 16 (pp. 315-320). IEEE.

41. Ljubic B, Hai AA, Stanojevic M, et al. Predicting complications of diabetes mellitus using advanced machine learning algorithms. J Am Med Inform Assoc. 2020 Sep 1;27(9):1343-51.

42. Zhong Z, Sun S, Weng J, et al. Machine learning algorithms identifying the risk of new-onset ACS in patients with type 2 diabetes mellitus: A retrospective cohort study. Front Public Health. 2022;10.

43. Lyu J, Li Z, Wei H, et al. A potent risk model for predicting new-onset acute coronary syndrome in patients with type 2 diabetes mellitus in Northwest China. Acta Diabetol. 2020 Jun;57(6):705-13.

44. Segar MW, Patel KV, Vaduganathan M, et al. Development and validation of optimal phenomapping methods to estimate long-term atherosclerotic cardiovascular disease risk in patients with type 2 diabetes. Diabetologia. 2021 Jul;64(7):1583-94.

45. Shi R, Wu B, Niu Z, Sun H, Hu F. Nomogram based on risk factors for type 2 diabetes mellitus patients with coronary heart disease. Diabetes Metab Syndr Obes. 2020;13:5025.

46. El Sanadi CE, Pantalone KM, Ji X, Kattan MW. Development and Internal Validation of A Prediction Tool To Assist Clinicians Selecting Second-Line Therapy Following Metformin Monotherapy For Type 2 Diabetes. Endocr Pract. 2021 Apr 1;27(4):334-41.

47. Basu S, Sussman JB, Berkowitz SA, Hayward RA, Yudkin JS. Development and validation of Risk Equations for Complications Of type 2 Diabetes (RECODe) using individual participant data from 5 randomized trials. Lancet Diabetes Endocrinol. 2017 Oct 1;5(10):788-98.

48. Hu WS, Lin CL. Use of the progression of adapted Diabetes Complications Severity Index to predict acute coronary syndrome, ischemic stroke, and mortality in Asian patients with type 2 diabetes mellitus: A nationwide cohort investigation. Clin Cardiol. 2018 Aug;41(8):1038-43.

49. Xiao S, Dong Y, Huang B, Jiang X. Predictive nomogram for coronary heart disease in patients with type 2 diabetes mellitus. Front Cardiovasc Med. 2022;9.

50. Koteliukh M. Predictive model for recurrent myocardial infarction in patients with type 2 diabetes mellitus. Med Sci. 2022.

51. Kazemian P, Wexler DJ, Fields NF, Parker RA, Zheng A, Walensky RP. Development and validation of PREDICT-DM: a new microsimulation model to project and evaluate complications and treatments of type 2 diabetes mellitus. Diabetes Technol Ther. 2019 Jun 1;21(6):344-55.

52. Lee SH, Han K, Kim HS, Cho JH, Yoon KH, Kim MK. Predicting the development of myocardial infarction in middle-aged adults with type 2 diabetes: a risk model generated from a nationwide population-based cohort study in Korea. Endocrinol Metab (Seoul). 2020 Sep;35(3):636.

53. Tam CH, Lim CK, Luk AO, et al. Development of genome-wide polygenic risk scores for lipid traits and clinical applications for dyslipidemia, subclinical atherosclerosis, and diabetes cardiovascular complications among East Asians. Genome Med. 2021 Dec;13(1):1-8.
54. Lithovius R, Antikainen AA, Mutter S, et al. Genetic Risk Score Enhances Coronary Artery Disease Risk Prediction in Individuals With Type 1 Diabetes. Diabetes Care. 2022 Mar;45(3):734-41.
55. Choi Y, Yang Y, Hwang BH, et al. Practical cardiovascular risk calculator for asymptomatic patients with type 2 diabetes mellitus: PRECISE-DM risk score. Clin Cardiol. 2020 Sep;43(9):1040-7.
56. Ferreira JP, Sharma A, Mehta C, et al. Multi-proteomic approach to predict specific cardiovascular events in patients with diabetes and myocardial infarction: findings from the EXAMINE trial. Clin Res Cardiol. 2021 Jul;110(7):1006-19.
57. Hirai H, Asahi K, Yamaguchi S, et al. New risk prediction model of coronary heart disease in participants with and without diabetes: assessments of the Framingham risk and Suita scores in 3-year longitudinal database in a Japanese population. Sci Rep. 2019 Feb 26;9(1):1-6.
58. Ye W, Ding X, Putnam N, et al. Development of clinical prediction models for renal and cardiovascular outcomes and mortality in patients with type 2 diabetes and chronic kidney disease using time-varying predictors. J Diabetes Complications. 2022 May 1;36(5):108180.
59. Hajar R. Risk factors for coronary artery disease: historical perspectives. Heart Views. 2017; 18(3): 109.
60. Quan J, Pang D, Li TK, et al. Risk prediction scores for mortality, cerebrovascular, and heart disease among Chinese people with type 2 diabetes. J Clin Endocrinol Metab. 2019 Dec;104(12):5823-30. Doi: 10.1210/jc.2019-00731.
61. Van der Heijden AA, Nijpels G, Badloe F, et al. Prediction models for development of retinopathy in people with type 2 diabetes: systematic review and external validation in a Dutch primary care setting. Diabetologia. 2020; 63(6):1110-1119.
62. Mukaka M, White SA, Terlouw DJ, Mwapasa V, Kalilani-Phiri L, Faragher EB. Is using multiple imputation better than complete case analysis for estimating a prevalence (risk) difference in randomized controlled trials when binary outcome observations are missing? Trials. 2016;17(1): 1-12.
63. Li X, Li F, Wang J, van Giessen A, Feenstra TL. Prediction of complications in health economic models of type 2 diabetes: a review of methods used. Acta Diabetol. 2023 Jul;60(7):861-79.
64. Caro JJ, Möller J, Karnon J, Stahl J, Ishak J. Discrete event simulation for health technology assessment. CRC press; 2015 Oct 16.
65. Roeters van Lennep JE, Westerveld HT, Erkelens DW, van der Wall EE. Risk factors for coronary heart disease: implications of gender. Cardiovasc Res. 2002; 53(3): 538-549.
66. Albus C. Psychological and social factors in coronary heart disease. Ann Med. 2010; 42(7): 487-494.
67. Kannel WB. Coronary heart disease risk factors in the elderly. Am J Geriatr Cardiol. 2002; 11(2): 101-107.
68. Hopkins PN, Williams RR. A survey of 246 suggested coronary risk factors. Atherosclerosis. 1981; 40(1): 1-52.
69. Collins GS, Dhiman P, Andaur Navarro CL et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11:e048008.
70. Wanner J, Herm LV, Janiesch C. How much is the black box? The value of explainability in machine learning models. ECIS 2020 Research-in-Progress Papers 2020; 85.
71. Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J. Explainable AI: A brief survey on history, research areas, approaches and challenges. InCCF international conference on natural language processing and Chinese computing 2019: 563-574.

# Appendices

**Appendix 1. The search strategy**

**Embase**
**Prediction**
("Validat$" OR "Predict$":ti OR "Rule$") OR ("Predict$" AND ("Outcome$" OR "Risk$" OR "Model$")) OR (("History" OR "Variable$" OR "Criteria" OR "Scor$" OR "Characteristic$" OR "Finding$" OR "Factor$") AND ("Predict$" OR "Model$" OR "Decision$" OR "Identif$" OR "Prognos$")) OR ("Decision$" AND ("Model$" OR "Clinical$")) OR ("Prognostic" AND ("History" OR "Variable$" OR "Criteria" OR "Scor$" OR "Characteristic$" OR "Finding$" OR "Factor$" OR "Model$")) OR ("Stratification" OR 'receiver operating characteristic'/exp OR "Discrimination" OR "Discriminate" OR "c-statistic" OR "c statistic" OR "Area under the curve" OR "AUC" OR "Calibration" OR "Indices" OR "Algorithm" OR "Multivariable")

**Coronary heart disease**
(("myocard*" OR "heart") NEAR/3 "infarct*"):ti,ab OR ("coronary heart" NEAR/3 ("disease" OR "occlus*" OR "angiogra*")):ti,ab OR 'acute coronary syndrome':ti,ab OR 'angina':ti,ab OR (("coronary artery") NEAR/3 ("disease" OR "obstruction" OR "atherosclerosis" OR "thrombosis")):ti,ab OR (("ischaemi*" OR "ischaemi*") NEAR/3 ("heart" OR "artery" OR "myocardial")):ti,ab OR 'ischemic heart disease'/exp

**Diabetes**
'diabetes mellitus'/exp OR "Diabet*":ti

**Exclusion of animal study**
"animal"/exp NOT "human"/exp

**Exclusion of systematic reviews**
("meta" OR "systematic review" OR "in vitro"):ti

**Publication type**
Limited to Articles and Conference paper

**Time duration**
2016 - 2021 (first round)
2021 - 2022 (second round)

**Appendix 1. Continued**

**PubMed**
**Prediction**
("Validat$" OR "Predict$" [ti] OR "Rule$") OR ("Predict$" AND ("Outcome$" OR "Risk$" OR "Model$")) OR (("History" OR "Variable$" OR "Criteria" OR "Scor$" OR "Characteristic$" OR "Finding$" OR "Factor$") AND ("Predict$" OR "Model$" OR "Decision$" OR "Identif$" OR "Prognos$")) OR ("Decision$" AND ("Model$" OR "Clinical$" OR "Logistic Models" [Mesh])) OR ("Prognostic" AND ("History" OR "Variable$" OR "Criteria" OR "Scor$" OR "Characteristic$" OR "Finding$" OR "Factor$" OR "Model$")) OR ("Stratification" OR "ROC Curve" [Mesh] OR "Discrimination" OR "Discriminate" OR "c-statistic" OR "c statistic" OR "Area under the curve" OR "AUC" OR "Calibration" OR "Indices" OR "Algorithm" OR "Multivariable")

**Coronary heart disease**
(("myocard*" [tiab] OR "heart" [tiab]) NEAR/3 infarct* [tiab]) OR ("coronary heart" [tiab] NEAR/3 ("disease" [tiab] OR "occlus*" [tiab] OR "angiogra*" [tiab])) OR 'acute coronary syndrome' [tiab] OR 'angina' [tiab] OR ("coronary artery" [tiab] NEAR/3 ("disease" [tiab] OR "obstruction" [tiab] OR "atherosclerosis" [tiab] OR "thrombosis" [tiab])) OR (("ischaemi*" [tiab] OR "ischaemi*" [tiab]) NEAR/3 ("heart" [tiab] OR "artery" [tiab] OR "myocardial" [tiab])) OR "Myocardial Ischemia"[Mesh]

**Diabetes**
"Diabetes Mellitus"[Mesh] OR "Diabet*" [ti]

**Exclusion of animal study**
"Animals"[Mesh] Not "Humans"[Mesh]

**Exclusion of systematic reviews**
"meta"[ti] OR "systematic review" [ti] OR "in vitro" [ti]

**Time duration**
January 2016 - May 2021 (first round)
June 2021 - December 2022 (second round)

**Appendix 1. Continued**

**Scopus**
**Prediction**
TITLE-ABS ("Predict$") OR ALL (("Validat$" OR "Rule$") OR ("Predict$" AND ("Outcome$" OR "Risk$" OR "Model$")) OR (("History" OR "Variable$" OR "Criteria" OR "Scor$" OR "Characteristic$" OR "Finding$" OR "Factor$") AND ("Predict$" OR "Model$" OR "Decision$" OR "Identif$" OR "Prognos$")) OR ("Decision$" AND ("Model$" OR "Clinical$")) OR ("Prognostic" AND ("History" OR "Variable$" OR "Criteria" OR "Scor$" OR "Characteristic$" OR "Finding$" OR "Factor$" OR "Model$")) OR ("Stratification" OR "Discrimination" OR "Discriminate" OR "c-statistic" OR "c statistic" OR "Area under the curve" OR "AUC" OR "Calibration" OR "Indices" OR "Algorithm" OR "Multivariable"))

**Coronary heart disease**
TITLE-ABS ((("myocard*" OR "heart") NEAR/3 "infarct*") OR ("coronary heart" NEAR/3 ("disease" OR "occlus*" OR "angiogra*")) OR 'acute coronary syndrome' OR 'angina' OR (("coronary artery") NEAR/3 ("disease" OR "obstruction" OR "atherosclerosis" OR "thrombosis")) OR (("ischaemi*" OR "ischaemi*") NEAR/3 ("heart" OR "artery" OR "myocardial")))

**Diabetes**
TITLE ("Diabet*")

**Systematic review**
TITLE ("meta" OR "systematic review" OR "in vitro")

**Time duration**
2016 - 2021 (first round)
2021 - 2022 (second round)

**Publication type**
Limited to Articles and Conference paper

**Web of Science**
**Prediction**
TI = ("Predict$") OR ALL = (("Validat$" OR "Rule$") OR ("Predict$" AND ("Outcome$" OR "Risk$" OR "Model$")) OR (("History" OR "Variable$" OR "Criteria" OR "Scor$" OR "Characteristic$" OR "Finding$" OR "Factor$") AND ("Predict$" OR "Model$" OR "Decision$" OR "Identif$" OR "Prognos$")) OR ("Decision$" AND ("Model$" OR "Clinical$")) OR ("Prognostic" AND ("History" OR "Variable$" OR "Criteria" OR "Scor$" OR "Characteristic$" OR "Finding$" OR "Factor$" OR "Model$")) OR ("Stratification" OR "Discrimination" OR "Discriminate" OR "c-statistic" OR "c statistic" OR "Area under the curve" OR "AUC" OR "Calibration" OR "Indices" OR "Algorithm" OR "Multivariable"))

**Coronary heart disease**
TS = ((("myocard*" OR "heart") NEAR/3 "infarct*") OR ("coronary heart" NEAR/3 ("disease" OR "occlus*" OR "angiogra*")) OR 'acute coronary syndrome' OR 'angina' OR (("coronary artery") NEAR/3 ("disease" OR "obstruction" OR "atherosclerosis" OR "thrombosis")) OR (("ischaemi*" OR "ischaemi*") NEAR/3 ("heart" OR "artery" OR "myocardial")))

**Diabetes**
TI = ("Diabet*")

**Systematic review**
TI = ("meta" OR "systematic review" OR "in vitro")

**Time duration**
January 1st 2016 - May 31st 2021 (first round)
June 1st 2021 - December 18th 2022 (second round)

**Appendix 1. Continued**

**IEEE**
**Diabetes AND Coronary heart disease**
("infarct*" OR "heart" OR "coronary" OR "artery" OR "angina" OR "ischaemi*" OR "ischaemi*") AND ("Abstract":"diabet*" OR "Document Title":"diabet*")

**Time duration**
2016 - 2021 (first round)
2021 - 2022 (second round)

**Publication type**
Limited to Journals and Conference paper

**Appendix 2. Signaling questions derived from Betts et al**

| No. | Signaling question to assess model applicability for the HTA purpose | Criteria | Critique mentioned by Betts et al, from which the signaling question was derived |
|---|---|---|---|
| a | Whether external validation of the model was performed in the same study? | / | Inappropriate geography/ generalizability |
| b | Whether the model was up-to-date (not outdated)? | The model was up-to-date if (1) the model was developed within ten years; and (2) the data sources used to develop the model were published within ten years. | Outdated |
| c | Whether the model includes all major CHD risk factors as candidate predictors or features? | The CHD risk factors/features refer to those mentioned by **Hajar R.** Risk factors for coronary artery disease: historical perspectives. Heart Views. 2017 Jul;18(3):109. More specifically, the risk factors/features include blood pressure, blood cholesterol levels, smoking, overweight or obesity, physical activity, diet and stress, age, sex , family history, and race. | Inappropriate covariates |

**Appendix 2. Continued**

| No. | Signaling question to assess model applicability for the HTA purpose | Criteria | Critique mentioned by Betts et al, from which the signaling question was derived |
|---|---|---|---|
| d | Whether the predictors were all likely to be reported as outcomes in clinical trials? | Only if the predictors in the model were usually reported as outcomes in clinical trials, the treatment effect can be modelled through effect on risk factors. Then the prediction model could be applied to the HTA modelling. A counter example is the photo. Some prediction models predict outcomes based on medical images created by ECG, CT, MRI. No trial will report how a treatment will change the photo. | Inappropriate covariates |
| e | Was an example given to illustrate how to calculate a risk using the model? | An example could be text, tables, figures, or online user interfaces that illustrate how a model could be used. For example, an artificial individual with assumed value on its characteristics is given to illustrate how an event risk is calculated. | Generalizability |
| f | Could the annual risk (or one-year probability of occurring an event) be directly calculated from the model? | The direct calculation indicates that, an equation or tool (e.g. an online user interface) is provided in the original study to calculate the annual risk of disease. | Generalizability |
| g | Could the annual risk (or one-year probability of occurring an event) be indirectly calculated from the model? | The indirect calculation indicates that, though equations or a tool for predicting the annual risk are not provided, users could calculate the risk, using evidence provided in the original study (e.g. using hazard functions to calculate the accumulated risk). | Generalizability |

**Appendix 3. Reference list of eligible studies**


1. Basu S, Sussman JB, Berkowitz SA, Hayward RA, Yudkin JS. Development and validation of Risk Equations for Complications Of type 2 Diabetes (RECODe) using individual participant data from randomised trials. Lancet Diabetes Endocrinol. 2017 Oct 1;5(10):788-98.
2. Choi Y, Yang Y, Hwang BH, et al. Practical cardiovascular risk calculator for asymptomatic patients with type 2 diabetes mellitus: PRECISE-DM risk score. Clin Cardiol. 2020 Sep;43(9):1040-7.
3. Fan R, Zhang N, Yang L, Ke J, Zhao D, Cui Q. AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus. Sci Rep. 2020 Sep 2;10(1):1-8.
4. Hirai H, Asahi K, Yamaguchi S, et al. New risk prediction model of coronary heart disease in participants with and without diabetes: assessments of the Framingham risk and Suita scores in 3-year longitudinal database in a Japanese population. Sci Rep. 2019 Feb 26;9(1):1-6.
5. Kazemian P, Wexler DJ, Fields NF, Parker RA, Zheng A, Walensky RP. Development and validation of PREDICT-DM: a new microsimulation model to project and evaluate complications and treatments of type 2 diabetes mellitus. Diabetes Technol Ther. 2019 Jun 1;21(6):344-55.
6. Kim E, Caraballo PJ, Castro MR, Pieczkiewicz DS, Simon GJ. Towards more accessible precision medicine: building a more transferable machine learning model to support prognostic decisions for micro-and macrovascular complications of type 2 diabetes mellitus. J Med Syst. 2019 Jul;43(7):1-2.
7. Lee SH, Han K, Kim HS, Cho JH, Yoon KH, Kim MK. Predicting the development of myocardial infarction in middle-aged adults with type 2 diabetes: a risk model generated from a nationwide population-based cohort study in Korea. Endocrinol Metab (Seoul). 2020 Sep;35(3):636.
8. Ljubic B, Hai AA, Stanojevic M, Diaz W, Polimac D, Pavlovski M, et al. Predicting complications of diabetes mellitus using advanced machine learning algorithms. J Am Med Inform Assoc. 2020 Sep 1;27(9):1343-51.
9. Longato E, Fadini GP, Sparacino G, Gubian L, Di Camillo B. Prediction of cardiovascular complications in diabetes from pharmacy administrative claims. In2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON) 2020 Jun 16 (pp. 315-320). IEEE.
10. Lyu J, Li Z, Wei H, Liu D, Chi X, Gong DW, et al. A potent risk model for predicting new-onset acute coronary syndrome in patients with type 2 diabetes mellitus in Northwest China. Acta Diabetol. 2020 Jun;57(6):705-13.
11. Quan J, Pang D, Li TK, Choi CH, Siu SC, Tang SY, et al. Risk prediction scores for mortality, cerebrovascular, and heart disease among Chinese people with type 2 diabetes. J Clin Endocrinol Metab. 2019 Dec;104(12):5823-30.
12. Segar MW, Patel KV, Vaduganathan M, Caughey MC, Jaeger BC, Basit M, et al. Development and validation of optimal phenomapping methods to estimate long-term atherosclerotic cardiovascular disease risk in patients with type 2 diabetes. Diabetologia. 2021 Jul;64(7):1583-94.
13. Shi R, Wu B, Niu Z, Sun H, Hu F. Nomogram based on risk factors for type 2 diabetes mellitus patients with coronary heart disease. Diabetes Metab Syndr Obes. 2020;13:5025.
14. El Sanadi CE, Pantalone KM, Ji X, Kattan MW. Development and Internal Validation of A Prediction Tool To Assist Clinicians Selecting Second-Line Therapy Following Metformin Monotherapy For Type 2 Diabetes. Endocr Pract. 2021 Apr 1;27(4):334-41.
15. Tam CH, Lim CK, Luk AO, Ng AC, Lee HM, Jiang G, et al. Development of genome-wide polygenic risk scores for lipid traits and clinical applications for dyslipidemia, subclinical atherosclerosis, and diabetes cardiovascular complications among East Asians. Genome Med. 2021 Dec;13(1):1-8.

16. Ferreira JP, Sharma A, Mehta C, Bakris G, Rossignol P, White WB, et al. Multi-proteomic approach to predict specific cardiovascular events in patients with diabetes and myocardial infarction: findings from the EXAMINE trial. Clin Res Cardiol. 2021 Jul;110(7):1006-19.

17. Hu WS, Lin CL. Use of the progression of adapted Diabetes Complications Severity Index to predict acute coronary syndrome, ischemic stroke, and mortality in Asian patients with type 2 diabetes mellitus: A nationwide cohort investigation. Clin Cardiol. 2018 Aug;41(8):1038-43.

18. Rådholm K, Chalmers J, Ohkuma T, Peters S, Poulter N, Hamet P, et al. Use of the waist-to-height ratio to predict cardiovascular risk in patients with diabetes: R esults from the ADVANCE-ON study. Diabetes Obes Metab. 2018 Aug;20(8):1903-10.

19. Koteliukh M. Predictive model for recurrent myocardial infarction in patients with type 2 diabetes mellitus. Med Sci. 2022.

20. Lithovius R, Antikainen AA, Mutter S, Valo E, Forsblom C, Harjutsalo V, et al. Genetic Risk Score Enhances Coronary Artery Disease Risk Prediction in Individuals With Type 1 Diabetes. Diabetes Care. 2022 Mar;45(3):734-41.

21. Xiao S, Dong Y, Huang B, Jiang X. Predictive nomogram for coronary heart disease in patients with type 2 diabetes mellitus. Front Cardiovasc Med. 2022;9.

22. Ye W, Ding X, Putnam N, Farej R, Singh R, Wang D, et al. Development of clinical prediction models for renal and cardiovascular outcomes and mortality in patients with type 2 diabetes and chronic kidney disease using time-varying predictors. J Diabetes Complications. 2022 May 1;36(5):108180.

23. Zhong Z, Sun S, Weng J, Zhang H, Lin H, Sun J, et al. Machine learning algorithms identifying the risk of new-onset ACS in patients with type 2 diabetes mellitus: A retrospective cohort study. Front Public Health. 2022;10.

24. Hayes AJ, Leal J, Gray AM, Holman RR, Clarke PM. UKPDS outcomes model 2: a new version of a model to simulate lifetime health outcomes of patients with type 2 diabetes mellitus using data from the 30 year United Kingdom Prospective Diabetes Study: UKPDS 82. Diabetologia. 2013 Sep;56:1925-33.

25. Nishimura K, Okamura T, Watanabe M, Nakai M, Takegami M, Higashiyama A, et al. Predicting coronary heart disease using risk factor categories for a Japanese urban population, and comparison with the framingham risk score: the suita study. J Atheroscler Thromb. 2014 Aug 26;21(8):784-98.

26. Piniés JA, González-Carril F, Arteagoitia JM, Irigoien I, Altzibar JM, Rodriguez-Murua JL, et al. Sentinel Practice Network of the Basque Country. Development of a prediction model for fatal and non-fatal coronary heart disease and cardiovascular disease in patients with newly diagnosed type 2 diabetes mellitus: the Basque Country Prospective Complications and Mortality Study risk engine (BASCORE). Diabetologia. 2014 Nov;57:2324-33.

**Appendix 4. Detailed characteristics of included 26 model studies**

| Study | Study design | Sample size | Region of the target population | Types of diabetes | Method of internal validation (Bootstraping & Split sample) |
|---|---|---|---|---|---|
| 1.Basu 2017 | Trial + Observational study | 9635 | US | Type-2 | 10-fold cross-validation |
| 2. Choi 2020 | Observational study | 933 | South Korea | Type-2 | 1000-fold cross-validation |
| 3. Fan 2020 | Observational study | 1273 | China | Type-2 | 5-fold cross-validation |
| 4. Hirai 2019 _ model 1 | Observational study | 2926 | Japan | All | NR |
| 4. Hirai 2019 _ model 2 | | | | | |
| 4. Hirai 2019 _ model 3 | | | | | |
| 5. Kazemian 2019 | Trial | 1800 | US | Type-2 | Sample split |
| 6. Kim 2019 | Observational study | 91429 | US | Type-2 | Bootstrapping with 500 iterations |
| 7. Lee 2020 | Observational study | 1,272,992 | South Korea | Type-2 | Sample split |
| 8. Ljubic 2020 | Observational study | 16439049 | US | Type-2 | Sample split & Cross-validation |
| 9. Longato 2020 | Observational study | 97466 | Italy | All | NR |
| 10. Lyu 2020 | Observational study | 456 | China | Type-2 | Sample split |
| 11. Quan 2019 | Observational study | 610647 | China & Singapore | Type-2 | Samples from 2 sources |
| 12. Segar 2021 | Trial + Observational study | 6466 | US | Type-2 | Samples from 2 sources |
| 13. Shi 2020 | Observational study | 3214 | China | Type-2 | Sample split |
| 14. El Sanadi 2021 | Observational study | 897 | US | Type-2 | Bootstrapping |

| External validation conducted? | Model type | Outcome of interest | Age of predicted individuals | Time cycle of prediction | Number of final predictors or top features |
|---|---|---|---|---|---|
| Yes | Cox proportional hazards model | MI | All | 10 years | 14 |
| Yes | Multivariate logistic regression model | Nonfatal MI | All | NR | 7 |
| Yes | Random forest based predictive model | CHD | All | NR | 8 |
| No | Cox proportional hazards model | CHD | All | 3 years | 4 |
| | Cox proportional hazards model | CHD | All | 3 years | 11 |
| | Cox proportional hazards model | CHD | All | 3 years | 14 |
| Yes | Cox regression model | MI | All | 10 years | 14 |
| Yes | Multi-Task Learning (MTL)-based model | CHD | All | NR | NR |
| No | Cox proportional hazards model | MI | 40-64 | 5 years | 12 |
| No | Recurrent neural network & gated recurrent unit (GRU) model | CHD | NR | 1 year | NR |
| No | Neural Network model | MI | NR | 3 years | NR |
| No | Multivariate logistic regression model | ACS | 25-85 | NR | 10 |
| Yes | Cox proportional hazards model | CHD | All | 5 years | 13 |
| Yes | Multivariable-adjusted Cox model with Finite mixture models | CHD | All | NR | 20 |
| Yes | Lead absolute shrinkage and selection operator (LASSO) regression model | CHD | 30-85 | NR | 8 |
| No | Fine and Gray's competing risk regression model | MI | 18-90 | NR | 26 |

**Appendix 4. Continued**

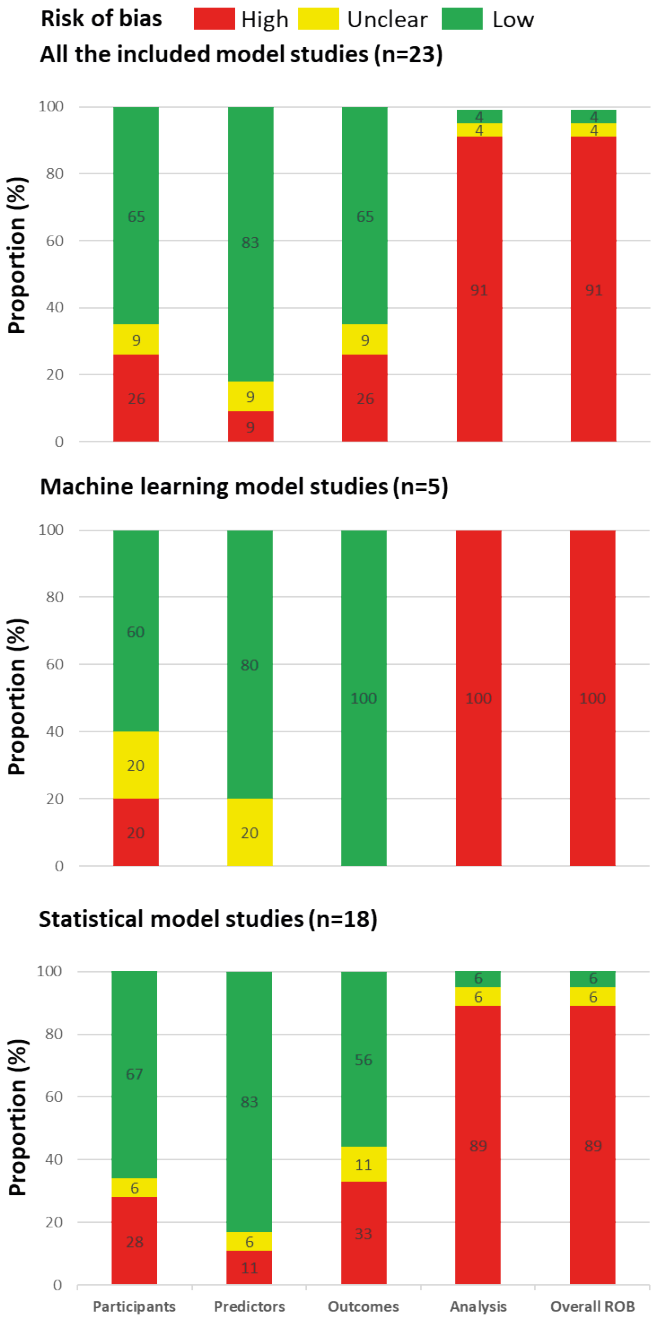
| Study | Study design | Sample size | Region of the target population | Types of diabetes | Method of internal validation (Bootstraping & Split sample) |
|---|---|---|---|---|---|
| 15. Tam 2021 | Observational study | 128305 | Japan | Type-2 | Samples from 2 sources |
| 16. Ferreira 2021 | Trial | 5131 | US | Type-2 | Bootstrapping |
| 17. Hu 2018 | Observational study | 84450 | China | Type-2 | NR |
| 18. Rådholm 2018 | Trial + Observational study | 11125 | 20 countries from Asia, Australia, Europe, and North America | Type-2 | Bootstrapping |
| 19. Koteliukh 2022 | Observational study | 74 | Ukraine | Type-2 | NR |
| 20. Lithovius 2022 | Observational study | 3295 | Finland | Type-1 | NR |
| 21. Xiao 2022 | Observational study | 560 | China | Type-2 | Sample split |
| 22. Ye 2022 | Observational study | 1432 | US | Type-2 | 10-fold cross-validation |
| 23. Zhong 2022 | Observational study | 521 | China | Type-2 | 5-fold cross-validation |
| 24. Hayes 2013_CHD | Observational study | 5102 | UK | Type-2 | Bootstrapping |
| 24. Hayes 2013_MI | | | | | |
| 25. Nishimura 2014_TC Suita Score | Observational study | 5521 | Japan | All | NR |
| 25. Nishimura 2014_LDL Suita Score | | | | | |
| 26. Piniés 2014 | Observational study | 777 | Spain | Type-2 | NR |

"CHD" indicates coronary heart disease; "MI", myocardial infarction; "ACS", acute coronary syndrome.

| External validation conducted? | Model type | Outcome of interest | Age of predicted individuals | Time cycle of prediction | Number of final predictors or top features |
|---|---|---|---|---|---|
| Yes | Logistic regression & polygenic risk score | CHD | NR | NR | NR |
| No | Cox regression model | MI | All | NR | 13 |
| Yes | Cox regression model | ACS | NR | NR | NR |
| No | Cox regression model | Nonfatal MI | NR | NR | NR |
| No | Generalized linear mixed model | MI | All | NR | 16 |
| No | Cox regression model | CHD | All | 5 years | NR |
| No | Logistic regression model | CHD | All | NR | 9 |
| No | Cox proportional-hazards model | MI | 30-74 | 1 year | 8 |
| No | K-nearest neighbor model | ACS | 30-100 | NR | 7 |
| No | Cox proportional hazards model | CHD | All | 1 year | 9 |
| | Cox proportional hazards model | MI | All | 1 year | 16 |
| No | Cox proportional hazards model | CHD | >=36 | 10 years | 8 |
| | Cox proportional hazards model | CHD | >=35 | 10 years | 8 |
| No | Cox regression model | CHD | All | 2 years | 5 |

**Appendix 5. PROBAST signaling questions for the 23 studies investigating models developed with statistical or machine learning techniques**

| Study No. | Study | DOMAIN 1: Participants | DOMAIN 2: Predictors | DOMAIN 3: Outcome | DOMAIN 4: Analysis |
|---|---|---|---|---|---|
| 1 | Basu 2017 | High | Low | High | High |
| 2 | Choi 2020 | Low | Low | Low | High |
| 3 | Fan 2020 | Unclear | Low | Low | High |
| 4 | Hirai 2019 | Low | Low | High | High |
| 5 | Kazemian 2019 | Unclear | Low | Low | Unclear |
| 6 | Kim 2019 | Low | Unclear | Low | High |
| 7 | Lee 2020 | High | Low | Low | High |
| 8 | Ljubic 2020 | High | Low | Low | High |
| 9 | Longato 2020 | Low | Low | Low | High |
| 10 | Lyu 2020 | High | High | High | High |
| 11 | Quan 2019 | Low | Low | Low | Low |
| 12 | Segar 2021 | Low | Unclear | Low | High |
| 13 | Shi 2020 | High | Low | High | High |
| 14 | El Sanadi 2021 | Low | Low | High | High |
| 15 | Tam 2021 | Low | Low | Low | High |
| 16 | Ferreira 2021 | Low | Low | Low | High |
| 17 | Hu 2018 | Low | Low | High | High |
| 18 | Rådholm 2018 | Low | High | Low | High |
| 19 | Koteliukh 2022 | High | Low | Unclear | High |
| 20 | Lithovius 2022 | Low | Low | Low | High |
| 21 | Xiao 2022 | Low | Low | Unclear | High |
| 22 | Ye 2022 | Low | Low | Low | High |
| 23 | Zhong 2022 | Low | Low | Low | High |




**Appendix 6. Overall risk of bias of included model studies (n=23)**

**Appendix 7. Applicability of the 30 models to the HTA settings**

| Study No. | Study | Signaling questions | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | a | b | c | d | e | f | g |
| 1 | Basu 2017 | PY | PN | PN | PY | Y | N | PN |
| 2 | Choi 2020 | PN | PN | N | PY | N | N | N |
| 3 | Fan 2020 | N | Y | N | PY | Y | N | N |
| 4 | Hirai 2019 _ model 1 | Y | PY | PN | PY | Y | N | N |
| 4 | Hirai 2019 _ model 2 | Y | PY | PN | PY | Y | N | N |
| 4 | Hirai 2019 _ model 3 | Y | PY | PN | PY | Y | N | N |
| 5 | Kazemian 2019 | Y | PN | NI | NI | N | N | PN |
| 6 | Kim 2019 | Y | PY | PN | PY | N | N | N |
| 7 | Lee 2020 | N | N | N | PY | Y | N | N |
| 8 | Ljubic 2020 | N | PN | NI | NI | N | N | N |
| 9 | Longato 2020 | N | PY | NI | NI | N | N | N |
| 10 | Lyu 2020 | N | Y | PN | PY | Y | NI | NI |
| 11 | Quan 2019 | PY | PN | PN | PY | Y | N | N |
| 12 | Segar 2021 | PY | PY | PY | PN | N | N | N |
| 13 | Shi 2020 | PN | Y | N | PY | Y | NI | NI |
| 14 | El Sanadi 2021 | N | Y | N | Y | Y | N | N |
| 15 | Tam 2021 | N | N | PN | PY | N | N | N |
| 16 | Ferreira 2021 | N | PY | PN | PN | N | N | N |
| 17 | Hu 2018 | N | PN | N | PY | N | N | N |
| 18 | Rådholm 2018 | N | PY | PN | PY | N | N | N |
| 19 | Koteliukh 2022 | N | Y | N | PN | N | N | N |
| 20 | Lithovius 2022 | N | PY | PN | Y | N | N | N |
| 21 | Xiao 2022 | N | Y | PN | PN | Y | NI | NI |
| 22 | Ye 2022 | N | Y | PN | Y | Y | Y | Y |
| 23 | Zhong 2022 | N | Y | PN | Y | Y | NI | NI |
| 24 | Hayes 2013_CHD | N | N | N | Y | Y | Y | Y |
| 24 | Hayes 2013_MI | N | N | N | Y | Y | Y | Y |
| 25 | Nishimura 2014_TC Suita Score | N | N | N | Y | Y | N | N |
| 25 | Nishimura 2014_LDL Suita Score | N | N | N | Y | Y | N | N |
| 26 | Piniés 2014 | Y | N | N | Y | Y | N | N |

"NI" indicates No information; "(P)Y", (Probably) Yes; "(P)N", (Probably) No.

6

# Part 3

## Statistical Methods for Incorporating Real-world Evidence into (Network) Meta-analyses

Chapter 7

# Comparison of Network Meta-analyses Investigating Efficacy of Diabetes Monitoring Systems with Insulin Delivery in Patients with Type-1 Diabetes, using Non-randomized Studies, Randomized-controlled Trials, or Both as Evidence

Li Jiu, Junfeng Wang, Jan-willem Versteeg, Jing Jin, Yingnan Deng, Konstantin Tashkov, Guenka Petrova, Olaf H Klungel, Aukje K Mantel-Teeuwisse, Wim G Goettsch

*Submitted*

# Abstract

## Background

Network meta-analyses (NMAs) have been conducted to investigate efficacy of diabetes monitoring systems (DMSs) combined with insulin delivery in patients with type-1 diabetes (T1DM), but previous NMAs only used randomized controlled trials (RCTs) as evidence. As statistical approaches that addressed quality concerns of non-randomized studies (NRSs) in NMAs emerged, we aimed to conduct parallel NMAs investigating DMSs with insulin delivery in T1DM patients, using NRSs, RCTs, or both as evidence, and investigated whether the estimated efficacy differed.

## Methods

RCTs were derived from Anthony et al. (2020), and NRSs were from the updated database search of Jiu et al. (2023). The target population was nonpregnant adult T1DM patients with at least 6 weeks of follow-up. Mean difference of HbA1c was the only outcome to be investigated, which could link more than three interventions via direct comparisons with a network map. We conducted NMAs with a Bayesian random-effects model, and downweighed NRSs using the power prior approach. We estimated and compared effect sizes and rankings, and tested whether assumptions related to missing data, model type, or NRSs' weight impacted the estimated efficacy.

## Results

Eighteen NRSs and 43 RCTs were included. RCTs belonged to two separate networks, and they were connected to one network, after NRSs were incorporated. The efficacy and rankings estimated from NRSs differed from those from RCTs, but results were not statistically significant. In contrast, results from RCTs and combined evidence were mostly similar. Additionally, changing the NRSs' weight relative to RCTs, especially for those with serious risk of bias, impacted the estimated efficacy greatly with statistical significance.

## Conclusions

NRSs, after being downweighed, could merge and extend the intervention networks of RCTs. Future research is needed to develop a good strategy to downweigh NRSs and RCTs based on risk of bias.

## Introduction

Type-1 diabetes (T1DM) is a chronic disease featured by insulin deficiency and resultant high blood glucose (1). Over the past three decades, T1DM has a global incidence of about 3% with 3-4% annual increase (2,3). T1DM also causes life-threatening complications with considerable financial and psychological burden (1, 4). For example, according the European National Health and Wellness Survey, adult T1DM patients reported significantly higher prevalence of comorbidities, such as hypertension, and lower health-related quality of life or work productivity loss, than those without diabetes (5). To manage T1DM, diabetes monitoring systems (DMSs), a portable medical device for monitoring blood glucose (6,7), are commonly used to administer dosage and dosage frequency of insulin, the cornerstone treatment for T1DM (8,9). Since DMSs are designed with different features, in terms of how blood glucose is measured or how an event is alarmed, DMSs have multiple categories, such as continuous glucose monitoring (CGM) and self-monitoring of blood glucose (SMBG) (10,11). Also, with DMS administration, insulin are normally delivered as cointerventions in two ways: continuous subcutaneous insulin infusion (CSII), i.e., insulin pump, and multiple daily injections (MDI) (11). Given the close association between DMSs and insulin delivery, the two concepts are sometimes integrated into one. For example, a closed-loop system or sensor-augmented pump (SAP) therapy comprises CGM, CSII, and an algorithm that responds to changes in glucose levels (12,13).



Given technical advances, the variety of DMSs, as well as the number of primary studies investigating DMSs' efficacy, has grown continuously in the past two decades (11). To compare the efficacy in a single analysis, several network meta-analyses (NMAs) have been conducted, and the most recent one was published by Anthony et al. (2020) (14). In their NMA, primary outcomes, including hemoglobine A1c (HbA1c), hypoglycemia rates, and quality of life, were compared among a maximum of 12 technologies (i.e. DMSs with insulin delivery) in nonpregnant adult T1DM patients. In addition, to ensure NMA validity, Anthony et al. (2020) operated on the exchangeability assumption, by including only randomized controlled trials (RCTs) as evidence. Exchangeability means all primary studies in a NMA are sufficiently similar, and that they measure the same underlying relative treatment effects (15).

While rigorously conducted, a NMA with only RCTs as evidence may not accurately predict results in real-world clinical practice, due to concerns on strict experimental settings or eligibility criteria for patients (16). Although these concerns could be addressed by combining RCTs and non-randomized studies (NRSs) in a NMA, the combination is rare in practice. One reason is that, NRSs' are considered of relatively

lower methodological quality and large between-study heterogeneity which pose serious challenges to the exchangeability assumption (17).

Recently, the development of novel statistical approaches has enabled the combination of RCTs and NRSs while adjusting for their differences in terms of methodological quality and between-study heterogeneity (16). One promising approach is power prior, which downweighs NRSs by functioning as priors of mean and variance of treatment effects in a Bayesian NMA using RCTs as evidence (18). However, in the case of DMSs, such novel approaches have not been applied, and a NMA including NRSs is still lacking.

Hence, the aim of our study was to conduct three parallel network meta-analyses, which used non-randomized studies, randomized-controlled trials, or both, to investigate efficacy of diabetes monitoring systems with insulin delivery as cointerventions in patients with type-1 diabetes. Also, we investigated whether the three analyses provided significantly different pooled estimates or ranking on the efficacy. This research was performed as part of the HTx project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825162 (19).

## Methods

### Sources of RCTs

All RCTs were derived from the NMA conducted by Anthony et al. (2020) (14). They identified 52 two-arm RCTs focusing on T1DM adult patients, which were published from inception to April, 2019, and they investigated efficacy of four outcomes, including HbA1c reduction (n=43), severe hypoglycemia (n=40), non-severe hypoglycemia (n=19), and quality of life (n=14). Quality of the RCTs were judged using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework. Since all RCTs and their relevant data (e.g. outcomes, patient demographics, etc.) were transparently reported, these RCTs could be directly synthesized for a NMA.

### Sources of NRSs

NRSs investigating efficacy of DMSs with insulin delivery in T1DM patients were partly derived from a pre-print review we conducted previously (20). In this review, 72 English-language retrospective studies focusing on any type of diabetes, which were published between January 2012 and March 2021, were identified. Since we planned to identify retrospective studies published from April 2021, and to include

prospective studies, we conducted an updated search of English-language articles published between 1st January, 2012 to 31st March, 2023. We searched four databases (i.e. Embase, PubMed, Web of Science, and Scopus) on 1st April, 2023, and screened the reference lists of studies considered eligible in the full-text review. The search strategy was the same as the one in our pre-print review.

### Eligibility criteria for NRSs

To enable comparison of the parallel NMAs, we followed the same eligibility criteria as Anthony et al. (2020) (14). A NRS was included if it (1) focused on nonpregnant adult patients (>=18 years old); (2) focused on T1DM; and (3) had a follow-up of at least 6 weeks. In our case, a NRS was defined as an observational study, and we excluded non-randomized experimental studies. In addition, we excluded NRSs with a single treatment group. Though methods to incorporate single-treatment-group studies into a NMA were available, such as matching, they were not recommended for studies with potentially large between-study heterogeneity (21).

### NRS identification & Data extraction

For the updated database search, one researcher (LJ) screened all titles and abstracts of identified records. The full-text of potentially eligible records, together with articles identified from our pre-print review, were reviewed independently by two researchers (LJ and JJ).



After identifying NRSs, two researchers (LJ and JV or YD) independently extracted outcome data, intervention information, and patient demographics (i.e. age, gender, duration of follow-up, and HbA1c at baseline). All discrepancies during study identification and data extraction were solved through discussion among the two, or by the third researcher (JW). Once missing data were identified, one researcher (LJ) contacted the study authors. Studies with missing outcome data were excluded.

Regarding interventions and outcomes, we followed the same definitions and categorizations as those from Anthony et al. (2020) (14). For example, DMSs were categorized as CGM, flash glucose monitoring (FGM), and SMBG. Also, a CGM with CSII that supported low-glucose suspend, glucose threshold alarms, or automated adjustment of insulin delivery was defined as an "integrated system".

### Selection of NRSs, given feasibility of conducting a NMA for a certain outcome

We planned to include all DMS-related outcomes for which a NMA is feasible. To assess the feasibility, we followed a process developed by Cope et al. (2014) (22). We

first judged whether a connected network is available, using the network map. If available, we judged whether the data were sufficient to explore the differences within or between direct treatment comparisons, which was often necessary for NRSs. The data sufficiency indicates whether meta-regression, subgroup analyses, or at least sensitivity analyses were available.

### Quality judgement of NRSs

Two researchers (LJ and JJ) independently judged quality of the eligible NRSs, using the Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I) checklist. NRSs identified from the previous review had already been judged using the ROBINs-I (20.) All discrepancies on quality judgement were solved through discussion.

### Data synthesis and analysis

All data syntheses and analyses were conducted, according to the Cochrane Handbook for Systematic Reviews of Interventions (23). For continuous outcomes, we estimated the mean difference and standard deviation(SD) between the baseline and end of follow-up. For binary outcomes, such as severe hypoglycemia, we estimated the odd ratio of event occurrence during the follow-up. If study covariates (e.g. age) or SDs (for continuous outcomes) at baseline or end of follow-up were missing, they were imputed with medians, within the same study type (i.e. RCTs or NRSs). To explore the impact of these missing SDs or covariates, we also imputed them with the maximum value in the sensitivity analysis. If SDs of mean difference (for continuous outcomes) were missing, they were imputed with the correlation coefficient approach cited by both the Cochrane handbook and Anthony et al. (2020) (14,23). We assumed that the correlation coefficient was 0.5, and explored its impact on efficacy, by replacing it with 0.1 or 0.9 in the sensitivity analysis.

We conducted three parallel NMAs, which used NRSs, RCTs, or the two as evidence. In the NMA with both as evidence, NRSs were downweighed using the power prior approach. We assumed that, compared to RCTs, a NRS with at least moderate risk of bias (RoB) , defined by ROBINS-I, had larger variance of prior distribution. More specifically, NRSs with low, moderate, serious, and critical RoB were assigned a variance one, two, four, and eight times relative to a RCT, respectively. To investigate the impact of assumed variances, we inflated or deflated NRS variances in four additional scenarios in the sensitivity analysis. In Scenario One, all NRSs were assigned the same variances as RCTs; in Scenario Two, the variances were two times for NRSs with moderate RoB, and eight times for serious-or-critical-RoB NRSs; In Scenario Three, the variances for NRSs with critical RoB were increased to 16 times,

compared to Scenario Two; In Scenario Four, the variances for moderate-RoB NRSs were increased to four times, compared to Scenario Three.

Since the approaches to downweigh non-randomized studies, including the power prior approach, were widely available in user-friendly R packages which adopted a Bayesian approach (24), we adopted the Bayesian approach for all the three NMAs. Though Anthony et al. (2020) had run a NMA for RCTs using the Frequentist approach, we repeated that NMA with the Bayesian approach, to minimize any potential impact of the approach on efficacy. All results were synthesized with random-effects models, and model impacts were tested with fixed-effect models in a sensitivity analysis. We ran Markov chain Monte Carlo (MCMC) simulations with 4 chains, 20000 samples, 5000 burn-ins, and without thinning. The MCMC convergence was tested using the Brooks-Gelman-Rubin method (Rhat), and was considered acceptable if Rhat was less than 1.1 (25). To test heterogeneity of studies within each comparison, we used the I-squared statistics (acceptable if > 40%) with 95% credible intervals (26). To test consistency among the studies, we used the node-splitting approach to assess whether direct and indirect evidence on a specific node were in agreement (27). We ranked the paired interventions (i.e. DMS + insulin delivery), by calculating the surface under the cumulative ranking curve (SUCRA). All statistical analyses were conducted using the R software (Version 3.4.2), with the gemtc package.



The efficacy was presented in tables and forest plots with 95% credible intervals (CrIs). To compare efficacy obtained from the three NMAs, we narratively compared effect sizes and rankings in a table.

## Results

### Study selection, missing data, and study characteristics

The flow chart of selecting NRSs is shown in Figure 1. A total of 7124 records were identified after removing duplicates, and 6941 records were excluded after scanning titles and abstracts. Of the 183 records for the full-text review, 22 articles were considered eligible. Since only for the outcome HbA1c, the number of studies (n=18) was relatively sufficient to construct a network map and to conduct scenario analyses, the four articles only investigating other outcomes (e.g. severe hypoglycemia) were excluded. The reference list and characteristics of the included NRSs are shown in Appendix 1 and 2.

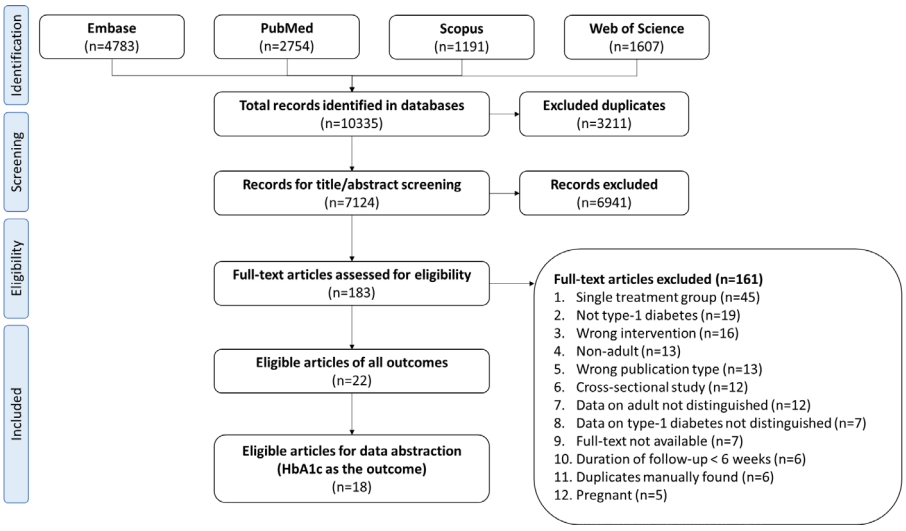



**Figure 1. Flow chart of NRS selection.**

More than half of these included NRSs had missing data for SDs of HbA1c value at baseline (n=6) or end of follow-up (n=9), or for SDs of the HbA1c reduction during the follow-up (i.e. mean difference) (n=14). All NRSs reported value for the four included covariates, but four provided covariates for the whole study population, rather than for each treatment group. No author responded to our request on missing data. Regarding covariates at the study level, NRSs had a longer diabetes duration (22 years vs. 19 years) than RCTs, but the medians in age (40.7 vs. 40.3 years), proportion of male patients (51% vs. 52%), and Hb1Ac at baseline (8.1% vs. 8.3%) were similar between the two. Regarding covariate distribution, the medians varied among the NRSs (also see Appendix 3). More specifically, the mean age, male proportion, diabetes duration, and baseline HbA1c ranged between 24.9 and 56.5, 27.8 and 63.5, 13.8 and 29.3, and 7.4 and 8.75, respectively. For RCTs, according to Anthony et al. (2020) (14), all covariates met the assumption of transitivity. Regarding the sample size of each treatment group, NRSs had a median almost twice larger than the RCTs' (56 vs.29.5).

## Network maps

Figure 1 shows a network of 10 and 13 paired interventions for HbA1c reductions, using NRSs or both RCTs and NRSs as evidence. Nine interventions were included by both RCTs and NRSs, and FGM + CSII was the only intervention that were included only by NRSs. Also, RCTs had two disconnected networks (i.e. with or without mixed insulin delivery) for HbA1c. These two networks were connected after they were combined with NRSs. In addition, sample sizes of different interventions varied significantly.

For example, in the NRS-only network, the sample sizes ranged between 40 (Integrated system) and 6736 (SMBG+(CSII/MDI)). Further details on all direct and indirect comparisons and relevant sample sizes are shown in Appendix 4 and 5.

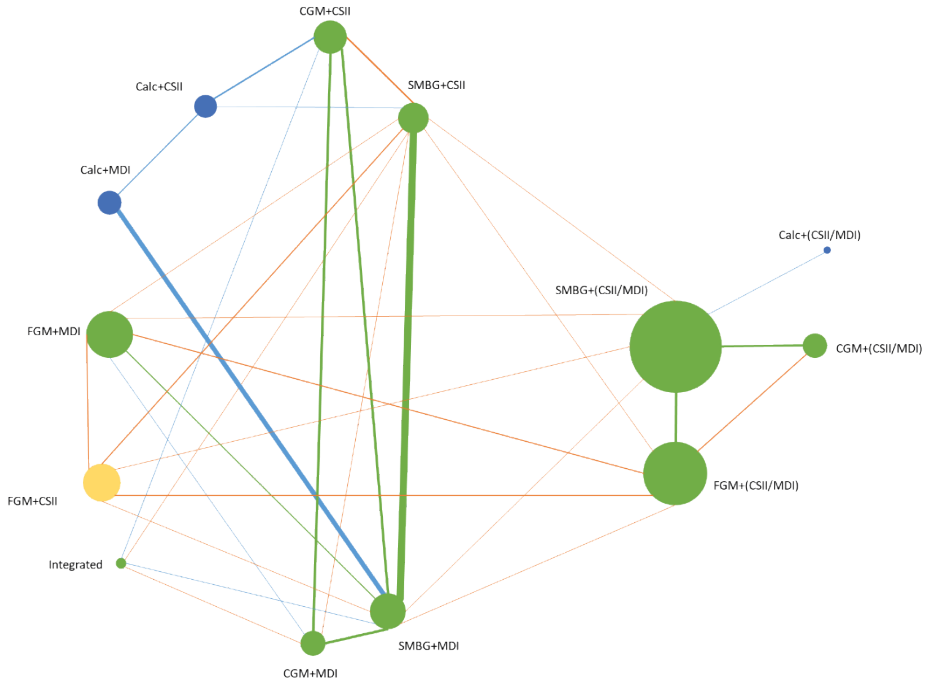



**Figure 2. Network map of diabetes monitoring systems combined with insulin delivery for HbA1c changes from baseline, using NRSs or both RCTs and NRSs as evidence.**
Blue, Yellow, and Green indicate that, an intervention (bubbles), or a pair of interventions (lines), was investigated by randomized controlled trials, non-randomized studies, and both, respectively. The size of each bubble and the width of each line is proportional to the number of patients receiving each intervention and the number of studies comparing each pair of interventions, respectively. CGM, continuous glucose monitoring; CSII, continuous subcutaneous insulin infusion; FGM, flash glucose monitoring; MDI, multiple daily injections; SMBG, self-monitoring of blood glucose; Calc, bolus calculators; (CSII/MDI), CSII and MDI both as within-study cointerventions

Regarding heterogeneity or consistency among primary studies, estimation was not possible in the NRS-only network, due to the relatively small number of included studies. In the NRS-and-RCT network, the heterogeneity was large among the both combined evidence (I-squared = 55%) and RCTs (I-squared = 64%) (see Appendix 6). According to the node-split approach (Appendix 7), the statistical consistency was present, because the differences among direct, indirect, and both evidence were not statistical significant for all paired comparisons.

### Comparison of NMAs, using RCTs, NRSs, or the both two as evidence

Effect sizes of each paired intervention, obtained using RCTs, NRSs, or both as evidence, are shown in Table 1. Given the existence of two disconnected RCT networks, we compared the interventions in two networks. Raw data extracted from each NRS are available in Appendix 8.

In summary, the mean differences of an intervention, analyzed from NRSs, differed significantly from those analyzed from RCTs. More specifically, in the first RCT disconnected network, the integrated system had the highest reduction of HbA1c from baseline (0.77) compared to other interventions, such as CGM+MDI. In contrast, according to NRSs, CGM+MDI was superior to other interventions, with a HbA1c reduction of 0.61 from baseline, while the integrated system only had a reduction of 0.05. Similarly, in the second RCT disconnected network, CGM+(CSII/MDI) was superior to FGM+(CSII/MDI), while the opposite result was observed in NRSs. In addition, the mean differences analyzed from both RCTs and NRSs were more similar to those obtained from RCTs than those from NRSs. For example, compared to SMBG+MDI, the HbA1c reduction of CGM+CSII was 0.68, 0.54, and 0.66, according to RCTs, NRSs, and the combined evidence, respectively. In addition, mean differences analyzed from NRSs incurred much larger uncertainty, in terms of 95% credible intervals, than RCTs, while the uncertainty was the smallest when analyzed from the combined evidence. Regarding the SUCRA ranking, RCTs and NRSs also showed significant difference. However, The SUCRA ranking was not completely consistent with results on the mean differences. For example, CGM+CSII ranked the first in the first RCT disconnected network, but its HbA1c reduction from baseline was less than the integrated system.

The sensitivity analysis shows that, for most interventions, the mean differences and 95% credible intervals were not sensitive to changed assumptions on imputed value for missing SDs or model types (i.e. random vs. fixed). In the NMA using NRSs as data sources, only the integrated system had a significant increase in HbA1c changes (from 0.05 to 0.19), as the missing SDs of HbA1c at baseline or end-of-follow-up were imputed with the maximum value rather than medians. In the NMA using RCTs or the combined data sources, CGM+MDI had a reduction of 0.14 in mean differences, when SDs of HbA1c at baseline or end-of-follow-up were imputed with the maximum value, or when a random-effect model was replaced with a fixed-effect model. While variations of mean differences were generally small in the sensitivity analyses, the relative rankings among interventions showing similar efficacy could change (e.g. the integrated system and CGM+CSII).

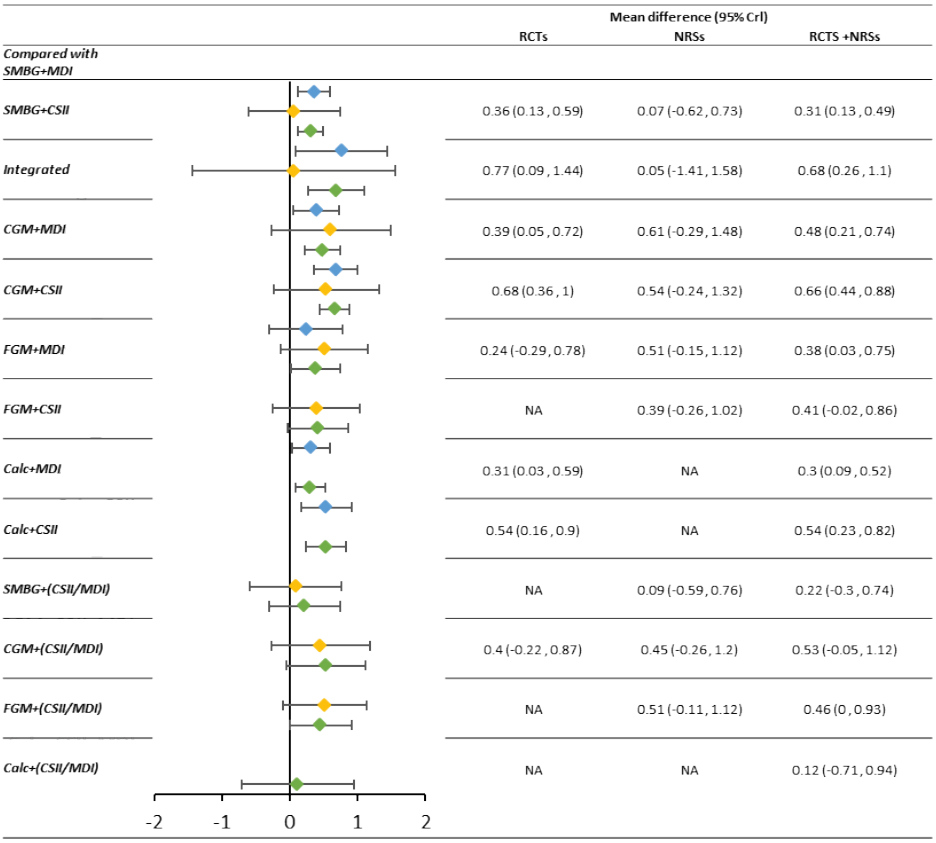
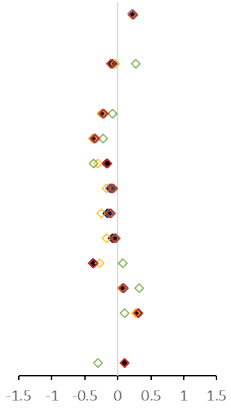



**Figure 3. Comparison of mean differences of HbA1c, with 95% credible intervals, between NMAs, using RCTs, NRSs, or the both two as evidence.**

Blue, Yellow, and Green indicates randomized controlled trials, non-randomized studies, and both, respectively. CGM, continuous glucose monitoring; CSII, continuous subcutaneous insulin infusion; FGM, flash glucose monitoring; MDI, multiple daily injections; SMBG, self-monitoring of blood glucose; Calc, bolus calculators; (CSII/MDI), CSII and MDI both as within-study cointerventions.




| | Base-case Scenario | Scenario One | Scenario Two | Scenario Three | Scenario Four |
|---|---|---|---|---|---|
| **Compared with Calc+CSII** | | | | | |
| Calc+MDI | 0.23 (-0.06 , 0.53) | 0.22 (-0.07 , 0.51) | 0.23 (-0.07 , 0.53) | 0.23 (-0.11 , 0.57) | 0.23 (-0.07 , 0.53) |
| **Compared with FGM+(CSII/MDI)** | | | | | |
| CGM+(CSII/MDI) | -0.07 (-0.49 , 0.34) | -0.04 (-0.35 , 0.27) | -0.1 (-0.53 , 0.34) | 0.28 (-0.1 , 0.65) | -0.1 (-0.54 , 0.34) |
| **Compared with SMBG+CSII** | | | | | |
| Calc+CSII | -0.23 (-0.55 , 0.09) | -0.25 (-0.56 , 0.05) | -0.22 (-0.55 , 0.1) | -0.08 (-0.66 , 0.5) | -0.22 (-0.54 , 0.11) |
| CGM+CSII | -0.35 (-0.61 , -0.09) | -0.38 (-0.59 , -0.16) | -0.35 (-0.61 , -0.08) | -0.22 (-0.59 , 0.15) | -0.34 (-0.61 , -0.07) |
| CGM+MDI | -0.17 (-0.47 , 0.14) | -0.3 (-0.55 , -0.04) | -0.16 (-0.47 , 0.15) | -0.36 (-0.69 , -0.04) | -0.16 (-0.47 , 0.15) |
| FGM+CSII | -0.11 (-0.55 , 0.33) | -0.17 (-0.53 , 0.19) | -0.11 (-0.57 , 0.34) | -0.07 (-0.61 , 0.46) | -0.08 (-0.57 , 0.42) |
| FGM+(CSII/MDI) | -0.16 (-0.63 , 0.32) | -0.25 (-0.63 , 0.12) | -0.13 (-0.66 , 0.39) | -0.1 (-0.69 , 0.48) | -0.1 (-0.65 , 0.44) |
| FGM+MDI | -0.08 (-0.46 , 0.3) | -0.18 (-0.51 , 0.15) | -0.05 (-0.45 , 0.34) | -0.03 (-0.47 , 0.42) | -0.04 (-0.45 , 0.37) |
| Integrated | -0.38 (-0.82 , 0.07) | -0.28 (-0.64 , 0.09) | -0.37 (-0.82 , 0.08) | 0.08 (-0.62 , 0.78) | -0.36 (-0.81 , 0.09) |
| SMBG+(CSII/MDI) | 0.08 (-0.45 , 0.62) | 0.06 (-0.35 , 0.47) | 0.07 (-0.53 , 0.67) | 0.33 (0.12 , 0.53) | 0.1 (-0.51 , 0.7) |
| SMBG+MDI | 0.31 (0.13 , 0.49) | 0.27 (0.1 , 0.43) | 0.31 (0.13 , 0.5) | 0.1 (-0.59 , 0.79) | 0.32 (0.13 , 0.5) |
| **Compared with SMBG+(CSII/MDI)** | | | | | |
| Calc+(CSII/MDI) | 0.1 (-0.53 , 0.73) | 0.1 (-0.52 , 0.72) | 0.1 (-0.53 , 0.73) | -0.3 (-0.66 , 0.06) | 0.1 (-0.53 , 0.73) |

**Figure 4. Impact of assumed weights of non-randomized studies with different risk of bias.**
Base-case Scenario (black): NRSs with moderate RoB were assigned 2-times variance relative to RCTs; serious RoB, 4-times variance; critical RoB, 8-times variance. Scenario One (yellow), all NRSs were assigned the same weight as RCTs. Scenario Two (blue), NRSs with moderate RoB, 2-times variance; serious and critical RoB, 8-times variance. Scenario Three (green), NRSs with moderate RoB, 2-time variance; serious RoB, 8-time variance; critical RoB, 16-time variance. Scenario Four (red), NRSs with moderate RoB, 4-time variance; serious RoB, 8-time variance; critical RoB, 16-time variance. CGM, continuous glucose monitoring; CSII, continuous subcutaneous insulin infusion; FGM, flash glucose monitoring; MDI, multiple daily injections; SMBG, self-monitoring of blood glucose; Calc, bolus calculators; (CSII/MDI), CSII and MDI both as within-study cointerventions; RCTs: randomized-controlled trials.

### Scenario analysis: Impact of assumed weights of NRSs with different risk of bias on effect sizes

Figure 3 shows the impact of assumed weights of NRSs with different risk of bias, relative to RCTs, on effect sizes. In the first scenario, all NRSs with different RoB were treated the same as RCTs, while in the second, third, and fourth scenario, NRSs with serious, critical, and moderate RoB were assigned even a lower weight, as compared to the base case. The scenario analysis shows that, downweighing or not downweighing NRSs impacted the estimated efficacy significantly. As shown in the base-case and first scenario, three interventions, i.e., CGM+MDI, FGM+MDI, and the integrated system, had a changed estimated efficacy of at least 0.1. In addition, the estimated efficacy of diabetes monitoring systems with insulin delivery was sensitive to the weights of NRSs, especially for those with critical RoB. More specifically, as the weight of NRSs with critical RoB was downweighed from one-eighth (Scenario Two) to one sixteenth (Scenario Three) relative to RCTs, the estimated efficacy of half of the interventions changed more than 0.1, and for three interventions, even more than 0.3 (e.g. Calc+(CSII/MDI)).

## Discussion



We conducted three parallel network meta-analyses on the efficacy of diabetes monitoring systems combined with insulin delivery, in patients with type-1 diabetes, using evidence from randomized controlled trials, non-randomized studies, and both. The NRSs linked the two disconnected RCT networks into one, and extended the network by adding another intervention (i.e. FGM+CSII). The efficacy and rankings estimated from NRSs differed significantly from the RCTs', but the NRSs' results were not significant, and had large uncertainty. In contrast, results estimated from RCTs and combined evidence were mostly similar (e.g. CGM+CSII compared with SMBG + MDI), as the NMA was dominated by RCTs. Also, changing the NRSs' weight relative to RCTs, especially for those with serious risk of bias, defined by ROBINS-I, impacted the estimated efficacy significantly.

The findings on whether the RCTs and NRSs provided consistent pooled estimates differed among the previous studies across disease fields. Brockelmann et al. (2022) obtained pooled estimates of 129 pairs of interventions from any disease field in meta-analyses, and compared whether the estimates from RCTs and from cohort studies differed significantly (28). They found that, on average, pooled estimates from the two did not differ. Similarly, Hong et at. (2021) analyzed 74 pairs of pooled effect estimates from RCTs and observational studies, and detected significant difference in 20% pairs (29).

In contrast, with evidence from meta-analyses (or from RCTs and NRSs with large sample sizes), Hill et al. (2023) compared estimates of six drugs to treat COVID-19 infection, and found statistically significant evidence of benefit from NRSs that was not seen in RCTs (30). In our study, the relatively consistent results from RCTs and combined evidence could be explained by the small number of NRSs and the power prior approach which assigned NRSs a low weight. For interventions of which results remained inconsistent (e.g. FGM+MDI compared with SMBG+MDI), the small number of RCTs could be an explanation. Still, researchers may raise concerns on whether the efficacy of such interventions estimated from RCTs truly reflect the real-world efficacy. For design of future primary studies, regardless of study type, a higher priority may be given to such interventions when selecting a target intervention.

One implication of our study was that, though results obtained from NMAs using NRSs as evidence might not be precise or valid, NRSs provided additional information that might inform decision-making, after they were incorporated into NMAs. On the one hand, NRSs identified in our study merged the two disconnected RCT networks, and enabled comparison of a diabetes monitoring system with single (CSII or MDI) and mixed insulin delivery. Information on such comparison may be needed for a clinical or HTA decision-making in some patient subgroups. For example, mixed insulin delivery could be a good option in patients receiving large insulin doses or when enhanced insulin absorption is needed to control hyperglycemia (31). On the other hand, we could gain more confidence on efficacy of a DMS with insulin delivery, if RCTs and NRSs provided consistent estimates. For example, according to Anthony et al. (2020) (14) and our results, the FGM+MDI, which was investigated by 2 RCTs, was superior to SMBG+MDI, but the results had uncertainty (95% credible interval crossing the y axis). After two NRSs were incorporated, the 95% credible intervals shrank, and did not cross the y axis (Figure 3).

We also found that, downweighing NRSs based on RoB, complemented by scenario analyses to investigate relevant impacts, might be a good strategy to incorporate NRSs into a NMA. With such strategy, a NMA might more accurately predict intervention efficacy in real-world practice, while biased estimates caused by relatively high-RoB NRSs were somewhat mitigated. While excluding all high-RoB studies is common practice, leaving them in a NMA with a small weight (e.g. one-sixteenth relative to RCT) might be conducive to the full use of evidence, especially when evidence is scarce. However, it remains a challenge to make assumptions on weights for each RoB level, as our study showed a significant impact of such assumptions on estimated efficacy. To reach more consensus on downweighing NRSs with different RoB levels, quantitative

bias analyses (32,33) may be needed in the future to examine the association between estimated efficacy and the extent of bias.

### Study limitations

Our study still had limitations. One limitation was that the total number of included non-randomized studies was quite small. Consequently, only HbA1c was investigated, while other outcomes of interest, such as hypoglycemia, were excluded. Also, given the small number of studies, we could not test network heterogeneity or inconsistency for NRSs. Still, we could reasonably speculate that the network heterogeneity and inconsistency were high, due to the high heterogeneity in the network with combined evidence. Another limitation was that we did not exclude RCTs or NRSs with high RoB from NMAs, given the small proportion of studies with low-to-intermediate RoB. Seventeen of the 18 NRSs were graded as serious or critical RoB, using ROBINS-I (Appendix 2), while all RCTs were considered with high performance bias by Anthony et al (2020), using the GRADE framework (14). Although a rigid quality assessment of RCTs and NRSs helps improve the quality of secondary research, such as NMAs, it may not help inform decision-making in practice, if almost all RCTs were simply graded as high RoB without further distinguishing their quality. In addition to excluding all these studies for the decision-making, excluding or downweighing studies based on domain-specific RoB (e.g. confounding bias) might be an alternative strategy. However, the relevant impact needs to be quantified in future research.



## Conclusions

The estimated efficacy and ranking on diabetes monitoring systems combined with insulin delivery, in patients with type-1 diabetes, were mostly similar, between the network meta-analyses, using randomized-controlled trials or the combined evidence. NRSs, after being downweighed, could merge and extend the intervention networks of RCTs, and inform decision-making in clinical and HTA settings. Future research is needed to develop a good strategy to downweigh NRSs and RCTs, based on risk of bias.

## Author contribution

LJ scanned and judged eligibility of all relevant original articles, collected data, conducted statistical analysis, and wrote the manuscript. All co-authors contributed to this study and critically reviewed and approved the manuscript.

# References


1. DiMeglio LA, Evans-Molina C, Oram RA. Type 1 diabetes. Lancet 2018; 391(10138): 2449-2462.
2. Norris JM, Johnson RK, Stene LC. Type 1 diabetes—Early life origins and changing epidemiology. Lancet Diabetes Endocrinol 2020 Mar; 1;8(3):226-38.
3. Vehik K, Dabelea D. The changing epidemiology of type 1 diabetes: why is it going through the roof?. Diabetes Metab Rev 2011 Jan;27(1):3-13.
4. Sussman M, Benner J, Haller MJ, Rewers M, Griffiths R. Estimated lifetime economic burden of type 1 diabetes. Diabetes Technol Ther 2020; 22(2): 121-130.
5. Ryden A, Sörstadius E, Bergenheim K, et al. The humanistic burden of type 1 diabetes mellitus in Europe: examining health outcomes and the role of complications. PLoS One 2016; Nov 3;11(11):e0164977.
6. Whitmore C. Blood glucose monitoring: an overview. Br J Nurs 2012;21(10):583-7.
7. Bolla AS, Priefer R. Blood glucose monitoring-an overview of current and future non-invasive devices. Diabetes Metab Syndr 2020;14(5):739-51.
8. Lawal M. Management of diabetes mellitus in clinical practice. Br J Nurs 2008;17(17), 1106-1113.
9. Quianzon CC, Cheikh I. History of insulin. J Community Hosp Intern Med Perspect 2012; 2(2): 18701.
10. Kamusheva M, Tachkov K, Dimitrova M, et al. A systematic review of collective evidences investigating the effect of diabetes monitoring systems and their application in health care. Front Endocrinol 2021;12:636959.
11. Beck RW, Bergenstal RM, Laffel LM, Pickup JC. Advances in technology for management of type 1 diabetes. Lancet 2019, 394(10205): 1265-1273.
12. Boughton CK, Hovorka R. New closed-loop insulin systems. Diabetologia 2021, 64(5): 1007-1015.
13. Raj S, Chakera A. Sensor-augmented pump therapy: review of new NICE diagnostic guidance. Pract Diabetes Int 2016, 33(2): 47-48a.
14. Pease A, Lo C, Earnest A, Kiriakova V, Liew D, Zoungas S. The efficacy of technology in type 1 diabetes: a systematic review, network meta-analysis, and narrative synthesis. Diabetes Technol Ther 2020 May 1;22(5):411-21.
15. Chaimani A, Caldwell D, Li T, Higgins J, Salanti G. Chapter 11: Undertaking network meta-analyses. Available from: https://training.cochrane.org/handbook/current/chapter-11 [Accessed September 26, 2023].
16. Efthimiou O, Mavridis D, Debray TP, et al. Combining randomized and non-randomized evidence in network meta-analysis. Stat Med 2017; 36(8): 1210-1226.
17. Cameron C, Fireman B, Hutton B, et al. Network meta-analysis incorporating randomized controlled trials and non-randomized comparative cohort studies for assessing the safety and effectiveness of medical treatments: challenges and opportunities. Syst Rev 2015, 4(1): 1-8.
18. Jenkins DA, Hussein H, Martina R, Dequen-O'Byrne P, Abrams KR, Bujkiewicz S. Methods for the inclusion of real-world evidence in network meta-analysis. BMC Med Res Methodol 2021 Dec;21(1):1-9.
19. HTx: About HTx project. Available from: https://www.htx-h2020.eu/about-htx-project [Accessed September 25, 2023].
20. Jiu L, Wang J, Kamusheva M, et al. Methodological Quality of Retrospective Observational Studies Investigating Effects of Diabetes Monitoring Systems: a Systematic Review (DOI:10.21203/rs.3.rs-2223544/v1). 2023. Available from: https://www.researchsquare.com/article/rs-2223544/v1. [Accessed September 26,2023].

21. Leahy J, Thom H, Jansen JP, et al. Incorporating single-arm evidence into a network meta-analysis using aggregate level matching: assessing the impact. Stat Med. 2019 Jun 30;38(14):2505-23.

22. Cope S, Zhang J, Saletan S, Smiechowski B, Jansen JP, Schmid P. A process for assessing the feasibility of a network meta-analysis: a case study of everolimus in combination with hormonal therapy versus chemotherapy for advanced breast cancer. BMC Med 2014 Dec;12(1):1-7.

23. Higgins J, Thomas J. Cochrane Handbook for Systematic Reviews of Interventions. Available from: https://training.cochrane.org/handbook/current. [Accessed September 26, 2023].

24. Buccheri S, Franchina G, Romano S, et al. Clinical outcomes following intravascular imaging-guided versus coronary angiography–guided percutaneous coronary intervention with stent implantation: a systematic review and Bayesian network meta-analysis of 31 studies and 17,882 patients. JACC Cardiovasc Interv 2017 Dec 26;10(24):2488-98.

25. Vats D, Knudson C. Revisiting the gelman–rubin diagnostic. Stat Sci 2021; 36(4): 518-529.

26. Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. Intern Emerg Med 2017, 12(1), 103-111.

27. van Valkenhoef G, Dias S, Ades AE, Welton NJ. Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. Res Synth Methods 2016, 7(1), 80-93.

28. Bröckelmann N, Balduzzi S, Harms L, et al. Evaluating agreement between bodies of evidence from randomized controlled trials and cohort studies in medical research: a meta-epidemiological study. BMC Med 2022 May 11;20(1):174.

29. Hong YD, Jansen JP, Guerino J, et al. Comparative effectiveness and safety of pharmaceuticals assessed in observational studies compared with randomized controlled trials. BMC Med 2021 Dec;19(1):1-5.

30. Hill A, Mirchandani M. The dangers of non-randomized, observational studies: experience from the COVID-19 epidemic. J Antimicrob Chemother 2023 Feb;78(2):323-7.

31. Meneghini L, Sparrow-Bodenmiller J. Practical aspects and considerations when switching between continuous subcutaneous insulin infusion and multiple daily injections. Diabetes Technol Ther 2010 Jun 1;12(S1):S-109.

32. Petersen JM, Ranker LR, Barnard-Mayers R, MacLehose RF, Fox MP. A systematic review of quantitative bias analysis applied to epidemiological research. Int J Epidemiol 2021 Oct;50(5):1708-30.

33. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol 2014 Dec 1;43(6):1969-85.

# Appendices

**Appendix 1 .Reference list of included studies**


1. Anderson J, Attvall S, Sternemalm L, Pivodic A, Fahlén M, Hanås R, Ekeroth G, Lind M. Effect on glycemic control by short-and long-term use of continuous glucose monitoring in clinical practice. Journal of diabetes science and technology. 2011 Nov;5(6):1472-9.
2. Greve SV, Stilgren L. A pragmatic real-life study of flash glucose monitoring versus self-monitoring of blood glucose. Dan Med J. 2020 Jun 1;67(6):A07190404.
3. Hidefjäll P, Berg L. Patient controlled, off-label use of continuous glucose monitoring: real-world medical costs and effects of patient controlled sensor augmented pump therapy in adult patients type 1 diabetes. Journal of Diabetes Science and Technology. 2021 May;15(3):575-81.
4. Nana M, Moore SL, Ang E, Lee ZX, Bondugulapati LN. Flash glucose monitoring: Impact on markers of glycaemic control and patient-reported outcomes in individuals with type 1 diabetes mellitus in the real-world setting. Diabetes Research and Clinical Practice. 2019 Nov 1;157:107893.
5. Parkin CG, Graham C, Smolskis J. Continuous glucose monitoring use in type 1 diabetes: longitudinal analysis demonstrates meaningful improvements in HbA1c and reductions in health care utilization. Journal of diabetes science and technology. 2017 May;11(3):522-8.
6. Tsur A, Cahn A, Israel M, Feldhamer I, Hammerman A, Pollack R. Impact of flash glucose monitoring on glucose control and hospitalization in type 1 diabetes: a nationwide cohort study. Diabetes/metabolism research and reviews. 2021 Jan;37(1):e3355.
7. Viñals C, Quirós C, Giménez M, Conget I. Real-life management and effectiveness of insulin pump with or without continuous glucose monitoring in adults with type 1 diabetes. Diabetes Therapy. 2019 Jun;10:929-36.
8. Gil-Ibáñez MT, Aispuru GR. Cost-effectiveness analysis of glycaemic control of a glucose monitoring system (FreeStyle Libre®) for patients with type 1 diabetes in primary health care of Burgos. Enfermería Clínica (English Edition). 2020 Mar 1;30(2):82-8.
9. Moreno-Fernandez J, Pazos-Couselo M, González-Rodriguez M, Rozas P, Delgado M, Aguirre M, Garcia-Lopez JM. Clinical value of flash glucose monitoring in patients with type 1 diabetes treated with continuous subcutaneous insulin infusion. Endocrinología, Diabetes y Nutrición (English ed.). 2018 Dec 1;65(10):556-63.
10. Irace C, Cutruzzolà A, Nuzzi A, Assaloni R, Brunato B, Pitocco D, Tartaglione L, Di Molfetta S, Cignarelli A, Laviola L, Citro G. Clinical use of a 180-day implantable glucose sensor improves glycated haemoglobin and time in range in patients with type 1 diabetes. Diabetes, Obesity and Metabolism. 2020 Jul;22(7):1056-61.
11. Šoupal J, Petruželková L, Flekač M, Pelcl T, Matoulek M, Daňková M, Škrha J, Svačina Š, Prázný M. Comparison of different treatment modalities for type 1 diabetes, including sensor-augmented insulin regimens, in 52 weeks of follow-up: a COMISAIR study. Diabetes technology & therapeutics. 2016 Sep 1;18(9):532-8.
12. Šoupal J, Petruželková L, Grunberger G, Hásková A, Flekač M, Matoulek M, Mikeš O, Pelcl T, Škrha Jr J, Horová E, Škrha J. Glycemic outcomes in adults with T1D are impacted more by continuous glucose monitoring than by insulin delivery method: 3 years of follow-up from the COMISAIR study. Diabetes Care. 2020 Jan 1;43(1):37-43.
13. Préau Y, Armand M, Galie S, Schaepelynck P, Raccah D. Impact of switching from intermittently scanned to real-time continuous glucose monitoring systems in a type 1 diabetes patient French cohort: an observational study of clinical practices. Diabetes Technology & Therapeutics. 2021 Apr 1;23(4):259-67.

14. Maiorino MI, Bellastella G, Casciano O, Cirillo P, Simeon V, Chiodini P, Petrizzo M, Gicchino M, Romano O, Caruso P, Giugliano D. The effects of subcutaneous insulin infusion versus multiple insulin injections on glucose variability in young adults with type 1 diabetes: the 2-year follow-up of the observational METRO study. Diabetes technology & therapeutics. 2018 Feb 1;20(2):117-26.

15. Garg SK, Voelmle MK, Beatson CR, Miller HA, Crew LB, Freson BJ, Hazenfield RM. Use of continuous glucose monitoring in subjects with type 1 diabetes on multiple daily injections versus continuous subcutaneous insulin infusion therapy: a prospective 6-month study. Diabetes care. 2011 Mar 1;34(3):574-9.

16. de Vera-Gómez PR, Mateo-Rodríguez C, Vivas-López C, Serrano-Olmedo I, Méndez-Muros M, Morales-Portillo C, Jiménez MS, Hernández-Herrero C, Martínez-Brocca MA. Effectiveness of a flash glucose monitoring systems implementation program through a group and telematic educational intervention in adults with type 1 diabetes. Endocrinología, Diabetes y Nutrición (English ed.). 2022 Nov 1;69(9):657-68.

17. Quirós C, Alonso-Carril N, Rodríguez-Rodríguez S, Barahona MJ, Orois A, Simó-Servat A, Ramos M, Perea V. The Medtronic 780G advanced hybrid closed-loop system achieves and maintains good glycaemic control in type 1 diabetes adults despite previous treatment. Endocrinología, Diabetes y Nutrición. 2023 Feb 1;70(2):130-5.

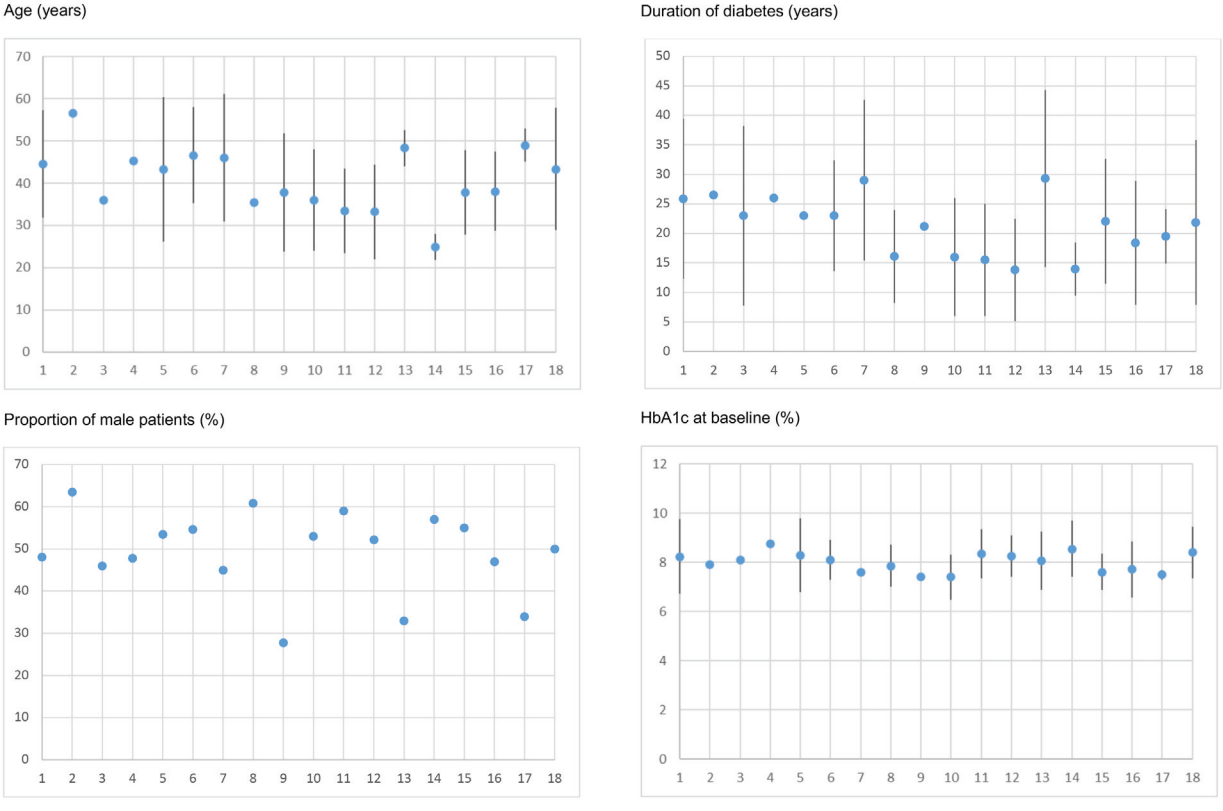
18. Brown RE, Chu L, Norman GJ, Abitbol A. Real-world glycaemic outcomes in adult persons with type 1 diabetes using a real-time continuous glucose monitor compared to an intermittently scanned glucose monitor: A retrospective observational study from the Canadian LMC diabetes registry (REAL-CGM-T1D). Diabetic Medicine. 2022 Nov;39(11):e14937.

**Appendix 2. Characteristics of included studies**

| Study | Intervention | ROB | Male(%) | Age | | Duration of diabetes | | HbA1c at baseline | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Mean | SD | Mean | SD |
| Anderson2011 | CGM+(CSII/MDI) | High | 55.9 | 44 | 10.3 | 26.4 | 11.8 | 8.79 | 1.6 |
| | SMBG+(CSII/MDI) | High | 47.5 | 44.6 | 15.7 | 25.8 | 15.5 | 8.19 | 1.4 |
| Greve 2020 | CGM+CSII | High | 46 | 36 | NR | 23 | NR | 8.1 | NR |
| | SMBG+CSII | High | 46 | 36 | NR | 23 | NR | 8.1 | NR |
| Hidefjäll 2021 | FGM+(CSII/MDI) | High | 47.8 | 45.3 | NR | 26 | 15.22 | 8.75 | NR |
| | SMBG+(CSII/MDI) | High | 47.8 | 45.3 | NR | 26 | 15.22 | 8.75 | NR |
| | FGM+CSII | High | 47.8 | 45.3 | NR | 26 | 15.22 | 8.75 | NR |
| | SMBG+CSII | High | 47.8 | 45.3 | NR | 26 | 15.22 | 8.75 | NR |
| | FGM+MDI | High | 47.8 | 45.3 | NR | 26 | 15.22 | 8.75 | NR |
| | SMBG+MDI | High | 47.8 | 45.3 | NR | 26 | 15.22 | 8.75 | NR |
| Nana 2019 | CGM+(CSII/MDI) | Low | 53.5 | 43.3 | NR | 23 | NR | 8.3 | NR |
| | SMBG+(CSII/MDI) | Low | 55.9 | 43.3 | NR | 23 | NR | 8.3 | NR |
| Parkin 2018 | FGM+(CSII/MDI) | High | 54.7 | 46.6 | 17.1 | 23 | NR | 8.1 | 1.5 |
| | FGM+CSII | High | 54.7 | 46.6 | 17.1 | 23 | NR | 8.1 | 1.5 |
| | FGM+MDI | High | 54.7 | 46.6 | 17.1 | 23 | NR | 8.1 | 1.5 |
| Tsur 2021 | CGM+CSII | High | 45 | 46 | 10 | 29 | 9.4 | 7.6 | 0.8 |
| | SMBG+CSII | High | 45 | 47 | 13 | 29 | 9.4 | 7.6 | 0.8 |
| Viñals 2019 | FGM+(CSII/MDI) | High | 60.9 | 35.4 | 15.1 | 16.1 | 13.6 | 7.86 | NR |
| | SMBG+(CSII/MDI) | High | 60.9 | 35.4 | 15.1 | 16.1 | 13.6 | 7.86 | NR |
| Gil-Ibáñez 2020 | FGM+CSII | High | 27.8 | 37.8 | NR | 21.2 | 8.6 | 7.4 | 0.7 |
| | SMBG+CSII | High | 27.8 | 38.6 | NR | 20.7 | 7.2 | 7.8 | 1 |
| Moreno-Fernandez 2018 | FGM+(CSII/MDI) | High | 62 | 56 | 15 | 27 | NR | 8 | NR |
| | SMBG+(CSII/MDI) | High | 65 | 57 | 13 | 26 | NR | 7.8 | NR |
| Irace 2020 | CGM+CSII | High | 54 | 36 | 12 | 16 | 10 | 7.3 | 0.92 |
| | CGM+MDI | High | 52 | 36 | 12 | 16 | 10 | 7.5 | 0.92 |
| Šoupal 2016 | CGM+MDI | High | 58 | 34 | 10 | 16 | 10 | 8.5 | 1.1 |
| | Integrated | High | 60 | 33 | 10 | 15 | 9 | 8.2 | 0.9 |
| Šoupal 2020 | CGM+MDI | High | 59 | 32.6 | 11.5 | 13.7 | 9.8 | 8.2 | 0.9 |
| | CGM+CSII | High | 50 | 32.3 | 9.9 | 14.6 | 7.8 | 8.2 | 0.9 |
| | SMBG+CSII | High | 48 | 33 | 9.3 | 13.4 | 8.4 | 8.3 | 0.8 |
| | SMBG+MDI | High | 52 | 35 | 15 | 13.5 | 8.8 | 8.3 | 0.8 |
| Préau 2021 | FGM+(CSII/MDI) | High | 33 | 48.3 | 4.3 | 29.3 | 15 | 8.07 | 1.18 |
| | CGM+(CSII/MDI) | High | 33 | 48.3 | 4.3 | 29.3 | 15 | 8.07 | 1.18 |

**Appendix 2. Continued**

| Study | Intervention | ROB | Male(%) | Age | | Duration of diabetes | | HbA1c at baseline | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Mean | SD | Mean | SD |
| Maiorino 2018 | CGM+CSII | High | 52 | 25.3 | 3.3 | 14.2 | 4.9 | 8.6 | 1.1 |
| | CGM+MDI | High | 62 | 24.5 | 2.9 | 13.7 | 4.1 | 8.5 | 1.2 |
| Garg 2011 | CGM+MDI | High | 50 | 39 | 11.35 | 22.2 | 10.14 | 7.62 | 0.68 |
| | CGM+CSII | High | 60 | 36.8 | 8.84 | 21.9 | 11.02 | 7.61 | 0.76 |
| de Vera-Gómez 2022 | FGM+(CSII/MDI) | High | 47 | 38.08 | 9.38 | 18.4 | 10.49 | 7.69 | 1.2 |
| | SMBG+(CSII/MDI) | High | 47 | 38.08 | 9.38 | 18.4 | 10.49 | 7.74 | 1.08 |
| Quirós 2023 | Integrated | High | 36 | 48 | 4.125 | 14 | 4.125 | 7.4 | 0.25 |
| | SMBG+CSII | High | 32 | 50 | 3.75 | 25 | 5.175 | 7.6 | 0.15 |
| Brown 2022 | CGM+(CSII/MDI) | High | 50 | 43.2 | 12.6 | 22.1 | 13.7 | 8.4 | 1.1 |
| | FGM+(CSII/MDI) | High | 50 | 43.5 | 16.7 | 21.6 | 14.2 | 8.4 | 1 |




**Appendix 3. Covariate distribution.**

**Appendix 4. Direct comparisons of interventions**

**Non-randomized studies**

| Comparison number | Intervention comparisons | Number of studies |
|---|---|---|
| 1 | SMBG+(CSII/MDI) vs CGM+(CSII/MDI) | 2 |
| 2 | SMBG+CSII vs CGM+CSII | 3 |
| 3 | FGM+CSII vs FGM+(CSII/MDI) | 2 |
| 4 | FGM+CSII vs FGM+MDI | 2 |
| 5 | FGM+MDI vs FGM+(CSII/MDI) | 2 |
| 6 | SMBG+(CSII/MDI) vs FGM+(CSII/MDI) | 4 |
| 7 | FGM+CSII vs SMBG+CSII | 2 |
| 8 | CGM+MDI vs CGM+CSII | 4 |
| 9 | CGM+MDI vs Integrated | 1 |
| 10 | CGM+MDI vs SMBG+CSII | 1 |
| 11 | SMBG+MDI vs CGM+MDI | 1 |
| 12 | FGM+(CSII/MDI) vs CGM+(CSII/MDI) | 2 |
| 13 | SMBG+MDI vs CGM+CSII | 1 |
| 14 | SMBG+MDI vs SMBG+CSII | 2 |
| 15 | Integrated vs SMBG+CSII | 1 |
| 16 | SMBG+CSII vs FGM+(CSII/MDI) | 1 |
| 17 | SMBG+MDI vs FGM+(CSII/MDI) | 1 |
| 18 | FGM+CSII vs SMBG+(CSII/MDI) | 1 |
| 19 | SMBG+CSII vs SMBG+(CSII/MDI) | 1 |
| 20 | SMBG+(CSII/MDI) vs FGM+MDI | 1 |
| 21 | SMBG+MDI vs SMBG+(CSII/MDI) | 1 |
| 22 | FGM+CSII vs SMBG+MDI | 1 |
| 23 | SMBG+CSII vs FGM+MDI | 1 |
| 24 | SMBG+MDI vs FGM+MDI | 1 |

**Non-randomized studies and randomized controlled trials**

| Comparison number | Intervention comparisons | Number of studies |
|---|---|---|
| 1 | SMBG+MDI vs SMBG+CSII | 13 |
| 2 | Calc+MDI vs SMBG+MDI | 9 |
| 3 | SMBG+MDI vs CGM+CSII | 5 |

**Appendix 4. Continued**

**Non-randomized studies and randomized controlled trials**

| Comparison number | Intervention comparisons | Number of studies |
|---|---|---|
| 4 | Calc+CSII vs CGM+CSII | 3 |
| 5 | SMBG+MDI vs CGM+MDI | 5 |
| 6 | CGM+MDI vs CGM+CSII | 5 |
| 7 | SMBG+(CSII/MDI) vs CGM+(CSII/MDI) | 4 |
| 8 | SMBG+(CSII/MDI) vs FGM+(CSII/MDI) | 5 |
| 9 | Calc+(CSII/MDI) vs SMBG+(CSII/MDI) | 1 |
| 10 | Calc+MDI vs Calc+CSII | 2 |
| 11 | Integrated vs CGM+CSII | 1 |
| 12 | SMBG+MDI vs Integrated | 1 |
| 13 | SMBG+MDI vs FGM+MDI | 2 |
| 14 | CGM+MDI vs FGM+MDI | 1 |
| 15 | Calc+CSII vs SMBG+CSII | 1 |
| 16 | SMBG+CSII vs CGM+CSII | 3 |
| 17 | FGM+MDI vs FGM+(CSII/MDI) | 2 |
| 18 | CGM+MDI vs Integrated | 1 |
| 19 | CGM+MDI vs SMBG+CSII | 1 |
| 20 | FGM+(CSII/MDI) vs CGM+(CSII/MDI) | 2 |
| 21 | Integrated vs SMBG+CSII | 1 |
| 22 | FGM+CSII vs FGM+(CSII/MDI) | 1 |
| 23 | SMBG+CSII vs FGM+(CSII/MDI) | 1 |
| 24 | SMBG+MDI vs FGM+(CSII/MDI) | 1 |
| 25 | FGM+CSII vs SMBG+(CSII/MDI) | 1 |
| 26 | SMBG+CSII vs SMBG+(CSII/MDI) | 1 |
| 27 | SMBG+(CSII/MDI) vs FGM+MDI | 1 |
| 28 | SMBG+MDI vs SMBG+(CSII/MDI) | 1 |
| 29 | FGM+CSII vs SMBG+CSII | 2 |
| 30 | FGM+CSII vs FGM+MDI | 2 |
| 31 | FGM+CSII vs SMBG+MDI | 1 |
| 32 | SMBG+CSII vs FGM+MDI | 1 |

**Appendix 5. Indirect comparisons of interventions**

**Non-randomized studies**

| Comparison number | Intervention comparisons |
|---|---|
| 1 | CGM+(CSII/MDI) vs CGM+CSII |
| 2 | CGM+(CSII/MDI) vs FGM+(CSII/MDI) |
| 3 | CGM+(CSII/MDI) vs FGM+MDI |
| 4 | CGM+(CSII/MDI) vs SMBG+CSII |
| 5 | CGM+(CSII/MDI) vs SMBG+MDI |
| 6 | CGM+(CSII/MDI) vs FGM+CSII |
| 7 | CGM+CSII vs FGM+(CSII/MDI) |
| 8 | CGM+CSII vs FGM+MDI |
| 9 | CGM+CSII vs SMBG+(CSII/MDI) |
| 10 | CGM+CSII vs SMBG+MDI |
| 11 | CGM+CSII vs FGM+CSII |
| 12 | FGM+(CSII/MDI) vs SMBG+CSII |
| 13 | FGM+MDI vs SMBG+(CSII/MDI) |
| 14 | SMBG+(CSII/MDI) vs SMBG+CSII |
| 15 | SMBG+(CSII/MDI) vs SMBG+MDI |
| 16 | SMBG+CSII vs SMBG+MDI |
| 17 | SMBG+MDI vs FGM+CSII |

**Non-randomized studies and randomized controlled trials**

| Comparison number | Intervention comparisons |
|---|---|
| 1 | CGM+(CSII/MDI) vs CGM+CSII |
| 2 | CGM+(CSII/MDI) vs FGM+(CSII/MDI) |
| 3 | CGM+(CSII/MDI) vs FGM+MDI |
| 4 | CGM+(CSII/MDI) vs SMBG+CSII |
| 5 | CGM+(CSII/MDI) vs Integrated |
| 6 | CGM+(CSII/MDI) vs CGM+MDI |
| 7 | CGM+(CSII/MDI) vs SMBG+MDI |
| 8 | CGM+(CSII/MDI) vs Calc+CSII |
| 9 | CGM+(CSII/MDI) vs Calc+(CSII/MDI) |
| 10 | CGM+(CSII/MDI) vs Calc+MDI |

**Appendix 5. Continued**

**Non-randomized studies and randomized controlled trials**

| Comparison number | Intervention comparisons |
|---|---|
| 11 | CGM+(CSII/MDI) vs FGM+CSII |
| 12 | CGM+CSII vs FGM+(CSII/MDI) |
| 13 | CGM+CSII vs FGM+MDI |
| 14 | CGM+CSII vs SMBG+(CSII/MDI) |
| 15 | CGM+CSII vs Calc+(CSII/MDI) |
| 16 | CGM+CSII vs Calc+MDI |
| 17 | CGM+CSII vs FGM+CSII |
| 18 | FGM+(CSII/MDI) vs SMBG+CSII |
| 19 | FGM+(CSII/MDI) vs Integrated |
| 20 | FGM+(CSII/MDI) vs CGM+MDI |
| 21 | FGM+(CSII/MDI) vs Calc+CSII |
| 22 | FGM+(CSII/MDI) vs Calc+(CSII/MDI) |
| 23 | FGM+(CSII/MDI) vs Calc+MDI |
| 24 | FGM+(CSII/MDI) vs FGM+CSII |
| 25 | FGM+MDI vs SMBG+(CSII/MDI) |
| 26 | FGM+MDI vs Integrated |
| 27 | FGM+MDI vs Calc+CSII |
| 28 | FGM+MDI vs Calc+(CSII/MDI) |
| 29 | FGM+MDI vs Calc+MDI |
| 30 | FGM+MDI vs FGM+CSII |
| 31 | SMBG+(CSII/MDI) vs SMBG+CSII |
| 32 | SMBG+(CSII/MDI) vs Integrated |
| 33 | SMBG+(CSII/MDI) vs CGM+MDI |
| 34 | SMBG+(CSII/MDI) vs SMBG+MDI |
| 35 | SMBG+(CSII/MDI) vs Calc+CSII |
| 36 | SMBG+(CSII/MDI) vs Calc+MDI |
| 37 | SMBG+CSII vs Integrated |
| 38 | SMBG+CSII vs CGM+MDI |
| 39 | SMBG+CSII vs Calc+(CSII/MDI) |
| 40 | SMBG+CSII vs Calc+MDI |
| 41 | Integrated vs CGM+MDI |

**Appendix 5. Continued**

**Non-randomized studies and randomized controlled trials**

| Comparison number | Intervention comparisons |
|---|---|
| 42 | Integrated vs Calc+CSII |
| 43 | Integrated vs Calc+(CSII/MDI) |
| 44 | Integrated vs Calc+MDI |
| 45 | Integrated vs FGM+CSII |
| 46 | CGM+MDI vs Calc+CSII |
| 47 | CGM+MDI vs Calc+(CSII/MDI) |
| 48 | CGM+MDI vs Calc+MDI |
| 49 | CGM+MDI vs FGM+CSII |
| 50 | SMBG+MDI vs Calc+CSII |
| 51 | SMBG+MDI vs Calc+(CSII/MDI) |
| 52 | SMBG+MDI vs FGM+CSII |
| 53 | Calc+CSII vs Calc+(CSII/MDI) |
| 54 | Calc+CSII vs FGM+CSII |
| 55 | Calc+(CSII/MDI) vs Calc+MDI |
| 56 | Calc+(CSII/MDI) vs FGM+CSII |
| 57 | Calc+MDI vs FGM+CSII |

**Appendix 6. Network heterogeneity**

**RCT (network 1)**

Per-comparison I-squared:

| | t1 | t2 | i2.pair | i2.cons | incons.p | incons.p |
|---|---|---|---|---|---|---|
| 1 | Calc+CSII | Calc+MDI | 0 | 77.79069 | 0.124245 | 0.476198 |
| 2 | Calc+CSII | CGM+CSII | 46.51302 | 19.21693 | 0.989707 | 0.747982 |
| 3 | Calc+CSII | SMBG+CSII | NA | 82.42327 | 0.087744 | 0.632722 |
| 4 | Calc+MDI | SMBG+MDI | 56.09916 | 63.70487 | 0.037695 | 0.552721 |
| 5 | CGM+CSII | CGM+MDI | NA | 90.31438 | 0.072223 | 0.722216 |
| 6 | CGM+CSII | Integrated | NA | 0 | 0.651372 | 0.935433 |
| 7 | CGM+CSII | SMBG+MDI | 74.21772 | 54.49806 | 0.180432 | 0.996065 |
| 8 | CGM+MDI | FGM+MDI | NA | 68.5137 | 0.180056 | 0.824182 |
| 9 | CGM+MDI | SMBG+MDI | 72.59192 | 63.77167 | 0.405991 | 0.889915 |
| 10 | FGM+MDI | SMBG+MDI | NA | 86.26791 | 0.131968 | 0.845673 |

**Appendix 6. Continued**

| **RCT (network 1)** | | | | | | |
|---|---|---|---|---|---|---|
| Per-comparison I-squared: | | | | | | |
| | t1 | t2 | i2.pair | i2.cons | incons.p | incons.p |
| 11 | Integrated | SMBG+MDI | NA | 0 | 0.906405 | 0.928383 |
| 12 | SMBG+CSII | SMBG+MDI | 67.74911 | 71.07069 | NA | 0.816157 |
| Global I-squared: | | | | | | |
| i2.pair | i2.cons | | | | | |
| 64.26361 | 69.92548 | | | | | |
| **RCT (network 2)** | | | | | | |
| Per-comparison I-squared: | | | | | | |
| | t1 | t2 | i2.pair | i2.cons | incons.p | |
| 1 | Calc+(CSII/MDI) | SMBG+(CSII/MDI) | NA | NA | NA | |
| 2 | CGM+(CSII/MDI) | SMBG+(CSII/MDI) | 63.55756 | 63.71811 | NA | |
| 3 | FGM+(CSII/MDI) | SMBG+(CSII/MDI) | NA | NA | NA | |
| Global I-squared: | | | | | | |
| i2.pair | i2.cons | | | | | |
| 63.57549 | 63.74205 | | | | | |
| **Both RCT and NRS** | | | | | | |
| Per-comparison I-squared: | | | | | | |
| | t1 | t2 | i2.pair | i2.cons | incons.p | |
| 1 | Calc_CSII | Calc_MDI | 0 | 81.02402 | 0.042587 | |
| 2 | Calc_CSII | CGM_CSII | 42.32592 | 11.97194 | 0.962313 | |
| 3 | Calc_CSII | SMBG_CSII | NA | 85.93439 | 0.031457 | |
| 4 | Calc_CSII_MDI | SMBG_CSII_MDI | NA | NA | NA | |
| 5 | Calc_MDI | SMBG_MDI | 59.78272 | 85.96035 | 6.69E-11 | |
| 6 | CGM_CSII | CGM_MDI | 0 | 39.15613 | 0.029378 | |
| 7 | CGM_CSII | Integrated | NA | 72.20842 | 0.261479 | |
| 8 | CGM_CSII | SMBG_CSII | 49.15238 | 39.13348 | 0.35326 | |
| 9 | CGM_CSII | SMBG_MDI | 58.77321 | 45.84222 | 0.142258 | |
| 10 | CGM_CSII_MDI | FGM_CSII_MDI | 0 | 0 | 0.633837 | |
| 11 | CGM_CSII_MDI | SMBG_CSII_MDI | 68.81884 | 63.28193 | 0.277025 | |
| 12 | CGM_MDI | FGM_MDI | NA | 55.53368 | 0.234481 | |

**Appendix 6. Continued**

**Both RCT and NRS**

Per-comparison I-squared:

| | t1 | t2 | i2.pair | i2.cons | incons.p |
|---|---|---|---|---|---|
| 13 | CGM_MDI | Integrated | NA | 0 | 0.62905 |
| 14 | CGM_MDI | SMBG_CSII | NA | 63.84136 | 0.147487 |
| 15 | CGM_MDI | SMBG_MDI | 63.52834 | 68.99992 | 0.038835 |
| 16 | FGM_CSII | FGM_CSII_MDI | 0 | 0 | NA |
| 17 | FGM_CSII | FGM_MDI | 0 | 0 | 0.421003 |
| 18 | FGM_CSII | SMBG_CSII | 0 | 58.70235 | 0.341638 |
| 19 | FGM_CSII | SMBG_CSII_MDI | NA | 0 | 0.907643 |
| 20 | FGM_CSII | SMBG_MDI | NA | 0 | 0.823591 |
| 21 | FGM_CSII_MDI | FGM_MDI | 18.14136 | 52.86468 | NA |
| 22 | FGM_CSII_MDI | SMBG_CSII | NA | 0 | 0.459865 |
| 23 | FGM_CSII_MDI | SMBG_CSII_MDI | 74.47075 | 67.82497 | 0.963438 |
| 24 | FGM_CSII_MDI | SMBG_MDI | NA | 0 | 0.551774 |
| 25 | FGM_MDI | SMBG_CSII | NA | 0 | 0.654871 |
| 26 | FGM_MDI | SMBG_CSII_MDI | NA | 15.0419 | 0.369389 |
| 27 | FGM_MDI | SMBG_MDI | 68.15415 | 91.2194 | 0.02198 |
| 28 | Integrated | SMBG_CSII | NA | 68.39913 | 0.136211 |
| 29 | Integrated | SMBG_MDI | NA | 0 | 0.520289 |
| 30 | SMBG_CSII | SMBG_CSII_MDI | NA | 0 | 0.90883 |
| 31 | SMBG_CSII | SMBG_MDI | 62.67129 | 59.64295 | 0.954611 |
| 32 | SMBG_CSII_MDI | SMBG_MDI | NA | 0 | 0.465163 |

Global I-squared:

| | i2.pair | i2.cons |
|---|---|---|
| 1 | 55.47647 | 59.16585 |

**Appendix 7. Study consistency, using the node-splitting approach**

| **Non-randomized studies** |
|---|
| Not applicable |

| | **Randomized controlled trials (Network 1)** | |
|---|---|---|
| | t1 | t2 |
| 1 | Calc+CSII | Calc+MDI |
| 2 | Calc+CSII | CGM+CSII |
| 3 | Calc+CSII | SMBG+CSII |
| 4 | Calc+MDI | SMBG+MDI |
| 5 | CGM+CSII | CGM+MDI |
| 6 | CGM+CSII | Integrated |
| 7 | CGM+CSII | SMBG+MDI |
| 8 | CGM+MDI | FGM+MDI |
| 9 | CGM+MDI | SMBG+MDI |
| 10 | FGM+MDI | SMBG+MDI |
| 11 | Integrated | SMBG+MDI |
| 12 | SMBG+CSII | SMBG+MDI |

| | **comparison** | **p.value** | **CrI** |
|---|---|---|---|
| 1 | d.Calc+CSII.Calc+MDI | 0.183225 | |
| 2 | -> direct | | 0.43 (-0.052, 0.91) |
| 3 | -> indirect | | -0.053 (-0.61, 0.50) |
| 4 | -> network | | 0.23 (-0.14, 0.59) |
| 5 | d.Calc+CSII.CGM+CSII | 0.946975 | |
| 6 | -> direct | | -0.15 (-0.60, 0.28) |
| 7 | -> indirect | | -0.13 (-0.76, 0.47) |
| 8 | -> network | | -0.14 (-0.51, 0.20) |
| 9 | d.Calc+CSII.SMBG+CSII | 0.0981 | |
| 10 | -> direct | | -0.40 (-1.2, 0.39) |
| 11 | -> indirect | | 0.37 (-0.093, 0.81) |
| 12 | -> network | | 0.18 (-0.24, 0.58) |
| 13 | d.Calc+MDI.SMBG+MDI | 0.188825 | |
| 14 | -> direct | | 0.40 (0.090, 0.69) |

**Appendix 7. Continued**

| | comparison | p.value | CrI |
|---|---|---|---|
| 15 | -> indirect | | -0.083 (-0.75, 0.58) |
| 16 | -> network | | 0.31 (0.029, 0.59) |
| 17 | d.CGM+CSII.CGM+MDI | 0.084475 | |
| 18 | -> direct | | -0.20 (-0.89, 0.49) |
| 19 | -> indirect | | 0.53 (0.049, 1.0) |
| 20 | -> network | | 0.29 (-0.12, 0.71) |
| 21 | d.CGM+CSII.Integrated | 0.915825 | |
| 22 | -> direct | | -0.100 (-0.79, 0.60) |
| 23 | -> indirect | | -0.0035 (-1.6, 1.6) |
| 24 | -> network | | -0.084 (-0.71, 0.54) |
| 25 | d.CGM+CSII.SMBG+MDI | 0.234225 | |
| 26 | -> direct | | 0.84 (0.43, 1.3) |
| 27 | -> indirect | | 0.46 (-0.019, 0.94) |
| 28 | -> network | | 0.68 (0.37, 1.0) |
| 29 | d.CGM+MDI.FGM+MDI | 0.229975 | |
| 30 | -> direct | | -0.20 (-1.0, 0.60) |
| 31 | -> indirect | | 0.46 (-0.31, 1.2) |
| 32 | -> network | | 0.15 (-0.41, 0.70) |
| 33 | d.CGM+MDI.SMBG+MDI | 0.5313 | |
| 34 | -> direct | | 0.32 (-0.079, 0.73) |
| 35 | -> indirect | | 0.56 (-0.086, 1.2) |
| 36 | -> network | | 0.39 (0.052, 0.72) |
| 37 | d.FGM+MDI.SMBG+MDI | 0.228325 | |
| 38 | -> direct | | 0.00026 (-0.67, 0.67) |
| 39 | -> indirect | | 0.66 (-0.22, 1.5) |
| 40 | -> network | | 0.24 (-0.29, 0.78) |
| 41 | d.Integrated.SMBG+MDI | 0.91765 | |
| 42 | -> direct | | 0.69 (-0.84, 2.3) |
| 43 | -> indirect | | 0.78 (0.018, 1.5) |
| 44 | -> network | | 0.77 (0.091, 1.4) |
| 45 | d.SMBG+CSII.SMBG+MDI | 0.102775 | |

**Appendix 7. Continued**

| | comparison | p.value | CrI |
|---|---|---|---|
| 46 | -> direct | | 0.31 (0.080, 0.54) |
| 47 | -> indirect | | 1.1 (0.18, 2.) |
| 48 | -> network | | 0.36 (0.13, 0.59) |

| **Randomized controlled trials (Network 2)** | | |
|---|---|---|
| Not applicable | | |
| **Both randomized controlled trials and non-randomized studies** | | |
| 1 | Calc+CSII | Calc+MDI |
| 2 | Calc+CSII | CGM+CSII |
| 3 | Calc+CSII | SMBG+CSII |
| 4 | Calc+MDI | SMBG+MDI |
| 5 | CGM+CSII | CGM+MDI |
| 6 | CGM+CSII | Integrated |
| 7 | CGM+CSII | SMBG+CSII |
| 8 | CGM+CSII | SMBG+MDI |
| 9 | CGM+MDI | FGM+MDI |
| 10 | CGM+MDI | SMBG+MDI |
| 11 | FGM+CSII | FGM+MDI |
| 12 | FGM+CSII | SMBG+CSII |
| 13 | FGM+CSII | SMBG+CSII+MDI |
| 14 | FGM+CSII | SMBG+MDI |
| 15 | FGM+CSII+MDI | SMBG+CSII |
| 16 | FGM+CSII+MDI | SMBG+MDI |
| 17 | FGM+MDI | SMBG+CSII |
| 18 | FGM+MDI | SMBG+CSII+MDI |
| 19 | FGM+MDI | SMBG+MDI |
| 20 | Integrated | SMBG+MDI |
| 21 | SMBG+CSII | SMBG+CSII+MDI |
| 22 | SMBG+CSII | SMBG+MDI |
| 23 | SMBG+CSII+MDI | SMBG+MDI |

**Appendix 7. Continued**

| | comparison | p.value | CrI |
|---|---|---|---|
| 1 | d.Calc+CSII.Calc+MDI | 0.114525 | |
| 2 | -> direct | | 0.43 (0.038, 0.82) |
| 3 | -> indirect | | -0.030 (-0.46, 0.40) |
| 4 | -> network | | 0.22 (-0.074, 0.51) |
| 5 | d.Calc+CSII.CGM+CSII | 0.8906 | |
| 6 | -> direct | | -0.14 (-0.50, 0.22) |
| 7 | -> indirect | | -0.096 (-0.54, 0.33) |
| 8 | -> network | | -0.12 (-0.41, 0.15) |
| 9 | d.Calc+CSII.SMBG+CSII | 0.043975 | |
| 10 | -> direct | | -0.40 (-1.1, 0.30) |
| 11 | -> indirect | | 0.40 (0.069, 0.71) |
| 12 | -> network | | 0.25 (-0.060, 0.55) |
| 13 | d.Calc+MDI.SMBG+MDI | 0.111125 | |
| 14 | -> direct | | 0.37 (0.14, 0.60) |
| 15 | -> indirect | | -0.085 (-0.62, 0.44) |
| 16 | -> network | | 0.30 (0.086, 0.51) |
| 17 | d.CGM+CSII.CGM+MDI | 0.05705 | |
| 18 | -> direct | | -0.079 (-0.36, 0.20) |
| 19 | -> indirect | | 0.35 (0.0053, 0.71) |
| 20 | -> network | | 0.078 (-0.15, 0.30) |
| 21 | d.CGM+CSII.Integrated | 0.3421 | |
| 22 | -> direct | | -0.10 (-0.66, 0.46) |
| 23 | -> indirect | | 0.24 (-0.22, 0.71) |
| 24 | -> network | | 0.095 (-0.26, 0.45) |
| 25 | d.CGM+CSII.SMBG+CSII | 0.231675 | |
| 26 | -> direct | | 0.54 (0.19, 0.90) |
| 27 | -> indirect | | 0.27 (-0.0070, 0.55) |
| 28 | -> network | | 0.38 (0.16, 0.59) |
| 29 | d.CGM+CSII.SMBG+MDI | 0.1623 | |
| 30 | -> direct | | 0.79 (0.50, 1.1) |
| 31 | -> indirect | | 0.52 (0.26, 0.77) |

**Appendix 7. Continued**

| | comparison | p.value | CrI |
|---|---|---|---|
| 32 | -> network | | 0.64 (0.45, 0.84) |
| 33 | d.CGM+CSII_MDI. FGM+(CSII/MDI) | 0.720575 | |
| 34 | -> direct | | 0.10 (-0.37, 0.57) |
| 35 | -> indirect | | -0.010 (-0.43, 0.41) |
| 36 | -> network | | 0.043 (-0.27, 0.35) |
| 37 | d.CGM+(CSII/MDI). SMBG+(CSII/MDI) | 0.726225 | |
| 38 | -> direct | | 0.33 (0.037, 0.63) |
| 39 | -> indirect | | 0.44 (-0.12, 1.0) |
| 40 | -> network | | 0.36 (0.100, 0.62) |
| 41 | d.CGM+MDI.FGM+MDI | 0.28215 | |
| 42 | -> direct | | -0.20 (-0.90, 0.49) |
| 43 | -> indirect | | 0.24 (-0.18, 0.65) |
| 44 | -> network | | 0.12 (-0.24, 0.48) |
| 45 | d.CGM+MDI.Integrated | 0.64625 | |
| 46 | -> direct | | 0.20 (-0.67, 1.1) |
| 47 | -> indirect | | -0.027 (-0.45, 0.41) |
| 48 | -> network | | 0.016 (-0.36, 0.40) |
| 49 | d.CGM+MDI.SMBG+CSII | 0.083 | |
| 50 | -> direct | | 0.90 (0.18, 1.6) |
| 51 | -> indirect | | 0.22 (-0.048, 0.49) |
| 52 | -> network | | 0.30 (0.042, 0.55) |
| 53 | d.CGM+MDI.SMBG+MDI | 0.091325 | |
| 54 | -> direct | | 0.38 (0.069, 0.68) |
| 55 | -> indirect | | 0.75 (0.43, 1.1) |
| 56 | -> network | | 0.57 (0.34, 0.79) |
| 57 | d.FGM+CSII.FGM+MDI | 0.303875 | |
| 58 | -> direct | | -0.13 (-0.58, 0.32) |
| 59 | -> indirect | | 0.30 (-0.42, 1.0) |
| 60 | -> network | | -0.0059 (-0.38, 0.36) |
| 61 | d.FGM+CSII.SMBG+CSII | 0.395125 | |

**Appendix 7. Continued**

| | comparison | p.value | CrI |
|---|---|---|---|
| 62 | -> direct | | 0.30 (-0.16, 0.77) |
| 63 | -> indirect | | -0.016 (-0.60, 0.59) |
| 64 | -> network | | 0.17 (-0.19, 0.53) |
| 65 | d.FGM+CSII.SMBG+(CSII/MDI) | 0.7049 | |
| 66 | -> direct | | 0.30 (-0.50, 1.1) |
| 67 | -> indirect | | 0.12 (-0.39, 0.63) |
| 68 | -> network | | 0.23 (-0.19, 0.65) |
| 69 | d.FGM+CSII.SMBG+MDI | 0.7231 | |
| 70 | -> direct | | 0.30 (-0.52, 1.1) |
| 71 | -> indirect | | 0.47 (0.053, 0.89) |
| 72 | -> network | | 0.44 (0.076, 0.80) |
| 73 | d.FGM+(CSII/MDI).SMBG+CSII | 0.18725 | |
| 74 | -> direct | | 0.70 (-0.10, 1.5) |
| 75 | -> indirect | | 0.082 (-0.37, 0.53) |
| 76 | -> network | | 0.25 (-0.12, 0.64) |
| 77 | d.FGM+(CSII/MDI).SMBG+(CSII/MDI) | 0.723125 | |
| 78 | -> direct | | 0.35 (0.045, 0.66) |
| 79 | -> indirect | | 0.23 (-0.32, 0.80) |
| 80 | -> network | | 0.32 (0.060, 0.58) |
| 81 | d.FGM+(CSII/MDI).SMBG+MDI | 0.482675 | |
| 82 | -> direct | | 0.70 (0.073, 1.3) |
| 83 | -> indirect | | 0.42 (-0.044, 0.90) |
| 84 | -> network | | 0.52 (0.16, 0.90) |
| 85 | d.FGM+MDI.SMBG+CSII | 0.40435 | |
| 86 | -> direct | | 0.50 (-0.33, 1.3) |
| 87 | -> indirect | | 0.11 (-0.27, 0.49) |
| 88 | -> network | | 0.17 (-0.15, 0.51) |
| 89 | d.FGM+MDI.SMBG+(CSII/MDI) | 0.2641 | |

**Appendix 7. Continued**

| | comparison | p.value | CrI |
|---|---|---|---|
| 90 | -> direct | | 0.50 (-0.12, 1.1) |
| 91 | -> indirect | | 0.052 (-0.46, 0.57) |
| 92 | -> network | | 0.23 (-0.16, 0.64) |
| 93 | d.FGM+MDI.SMBG+MDI | 0.073975 | |
| 94 | -> direct | | 0.20 (-0.19, 0.61) |
| 95 | -> indirect | | 0.77 (0.30, 1.3) |
| 96 | -> network | | 0.44 (0.13, 0.76) |
| 97 | d.Integrated.SMBG+CSII | 0.161075 | |
| 98 | -> direct | | -0.20 (-0.96, 0.56) |
| 99 | -> indirect | | 0.42 (0.0026, 0.82) |
| 100 | -> network | | 0.28 (-0.088, 0.64) |
| 101 | d.Integrated.SMBG+MDI | 0.57415 | |
| 102 | -> direct | | 0.70 (0.057, 1.3) |
| 103 | -> indirect | | 0.48 (0.046, 0.90) |
| 104 | -> network | | 0.55 (0.20, 0.90) |
| 105 | d.SMBG+CSII. SMBG+(CSII/MDI) | 0.90605 | |
| 106 | -> direct | | 0.0033 (-0.81, 0.82) |
| 107 | -> indirect | | -0.052 (-0.56, 0.46) |
| 108 | -> network | | 0.061 (-0.35, 0.47) |
| 109 | d.SMBG+CSII.SMBG+MDI | 0.527375 | |
| 110 | -> direct | | 0.25 (0.055, 0.44) |
| 111 | -> indirect | | 0.37 (0.024, 0.73) |
| 112 | -> network | | 0.27 (0.11, 0.44) |
| 113 | d.SMBG+(CSII/MDI). SMBG+MDI | 0.3661 | |
| 114 | -> direct | | -0.0035 (-0.62, 0.62) |
| 115 | -> indirect | | 0.37 (-0.16, 0.91) |
| 116 | -> network | | 0.21 (-0.19, 0.61) |

**Appendix 8. Data on HbA1c changes**

| Study | Intervention | Sample | Baseline mean | Baseline SD | End mean | End SD | Mean difference | SD |
|---|---|---|---|---|---|---|---|---|
| Anderson2011 | CGM+(CSII/MDI) | 34 | 8.79 | 1.85 | 8.19 | NR | -0.6 | NR |
| | SMBG+(CSII/MDI) | 408 | 8.19 | 1.35 | 8.19 | NR | 0 | NR |
| Greve 2020 | CGM+CSII | 187 | 8.3 | NR | 7.8 | NR | -0.5 | NR |
| | SMBG+CSII | 188 | 8.3 | NR | 8.3 | NR | 0 | NR |
| Hidefjäll 2021 | FGM+(CSII/MDI) | 83 | 8.7 | NR | 8 | NR | -0.7 | NR |
| | SMBG+(CSII/MDI) | 83 | 8.7 | NR | 8.7 | NR | 0 | NR |
| | FGM+CSII | 14 | 7.6 | NR | 7.3 | NR | -0.3 | NR |
| | SMBG+CSII | 14 | 7.6 | NR | 7.6 | NR | 0 | NR |
| | FGM+MDI | 63 | 8.6 | NR | 8.1 | NR | -0.5 | NR |
| | SMBG+MDI | 63 | 8.6 | NR | 8.6 | NR | 0 | NR |
| Nana 2019 | CGM+(CSII/MDI) | 164 | 8.24 | NR | 7.71 | NR | -0.53 | NR |
| | SMBG+(CSII/MDI) | 6006 | 8.27 | NR | 8.04 | NR | -0.23 | NR |
| Parkin 2018 | FGM+(CSII/MDI) | 2682 | 8.1 | 1.5 | 7.9 | 1.3 | -0.2 | NR |
| | FGM+CSII | 1118 | 7.9 | 1.2 | 7.8 | 1.2 | -0.1 | NR |
| | FGM+MDI | 1564 | 8.2 | 1.6 | 8 | 1.4 | -0.2 | NR |
| Tsur 2021 | CGM+CSII | 40 | 7.6 | 0.8 | 7.4 | 0.7 | -0.2 | NR |
| | SMBG+CSII | 120 | 7.6 | 0.8 | 7.7 | 0.9 | 0.1 | NR |
| Viñals 2019 | FGM+(CSII/MDI) | 23 | 7.86 | NR | 7.47 | NR | -0.39 | NR |
| | SMBG+(CSII/MDI) | 23 | 7.86 | NR | 7.86 | NR | 0 | NR |
| Gil-Ibáñez 2020 | FGM+CSII | 18 | 7.4 | 0.7 | 7.1 | 0.7 | -0.3 | 0.2 |
| | SMBG+CSII | 18 | 7.8 | 1 | 7.8 | 1 | 0 | 0.18 |
| Moreno-Fernandez 2018 | FGM+(CSII/MDI) | 128 | 8 | NR | 7.6 | NR | -0.4 | 1.11 |
| | SMBG+(CSII/MDI) | 128 | 7.9 | NR | 8 | NR | 0.1 | 1.11 |

**Appendix 8. Continued**

| Study | Intervention | Sample | Baseline mean | Baseline SD | End mean | End SD | Mean difference | SD |
|---|---|---|---|---|---|---|---|---|
| Irace 2020 | CGM+CSII | 56 | 7.3 | NR | 6.9 | NR | -0.4 | NR |
| | CGM+MDI | 44 | 7.5 | NR | 7 | NR | -0.5 | NR |
| Šoupal 2016 | CGM+MDI | 12 | 8.5 | 1.1 | 7.2 | 0.8 | -1.3 | NR |
| | Integrated | 15 | 8.2 | 0.9 | 7.1 | 0.9 | -1.1 | NR |
| Šoupal 2020 | CGM+MDI | 22 | 8.2 | 0.9 | 7 | NR | -1.2 | NR |
| | CGM+CSII | 26 | 8.2 | 0.9 | 6.9 | NR | -1.3 | NR |
| | SMBG+CSII | 25 | 8.3 | 0.8 | 7.7 | NR | -0.3 | NR |
| | SMBG+MDI | 21 | 8.3 | 0.8 | 8 | NR | -0.6 | NR |
| Préau 2021 | FGM+(CSII/MDI) | 18 | 8.07 | 1.18 | 8.07 | 1.18 | 0 | NR |
| | CGM+(CSII/MDI) | 18 | 8.07 | 1.18 | 8.19 | 1.11 | 0.12 | 0.75 |
| Maiorino 2018 | CGM+CSII | 98 | 8.6 | 1.1 | 8.1 | 1 | -0.41 | 1.07 |
| | CGM+MDI | 125 | 8.5 | 1.2 | 8.1 | 1.3 | -0.42 | 1 |
| Garg 2011 | CGM+MDI | 30 | 7.62 | 0.68 | 7.78 | 1.03 | -0.16 | NR |
| | CGM+CSII | 30 | 7.61 | 0.76 | 7.59 | 0.91 | -0.02 | NR |
| de Vera-Gómez 2022 | FGM+(CSII/MDI) | 88 | 7.69 | 1.2 | 7.17 | 1 | -0.45 | 0.0875 |
| | SMBG+(CSII/MDI) | 88 | 7.74 | 1.08 | 7.69 | 1.2 | -0.05 | NR |
| Quirós 2023 | Integrated | 25 | 7.4 | NR | 7 | NR | -0.4 | NR |
| | SMBG+CSII | 25 | 7.6 | NR | 7 | NR | -0.6 | NR |
| Brown 2022 | CGM+(CSII/MDI) | 143 | 8.4 | 1.1 | 7.7 | 1 | -0.7 | NR |
| | FGM+(CSII/MDI) | 143 | 8.4 | 1 | 7.9 | 1.1 | -0.5 | NR |

7

Chapter 8

# Approaches to Synthesizing Evidence from Randomized Controlled Trials and Non-randomized Studies in Meta-analyses: Application of the Crossnma Package to the Cases of Myelodysplastic Syndromes and Diabetes

Li Jiu, Junfeng Wang, Olaf H Klungel, Aukje K Mantel-Teeuwisse, Wim G Goettsch

*Manuscript in preparation*

# Abstract

## Background

Statistical approaches supporting synthesis of randomized controlled trials (RCTs) and non-randomized studies (NRSs), e.g. naïve pooling, power prior, and hierarchical modelling, in a meta-analysis have emerged, but they have not been widely applied or compared in practice. A newly developed tool, the Crossnma package, which enables comparison of the three approaches in one R software and supports relevant uncertainty analyses, may solve the practical difficulties researchers have faced. We aimed to apply this package to cases on myelodysplastic syndromes (MDS) or diabetes, to explore whether the effect estimates obtained from the three approaches were consistent.

## Methods

The naïve pooling, power prior, and hierarchical modelling approach were applied in four cases: two case studies (CS) on MDS (CS1 and CS2) and two (CS3 and CS4) on diabetes. The cases differed in proportions of RCTs and number of primary studies. We also conducted sensitivity analyses to investigate whether covariates (i.e. age and duration of follow-up) or the NRSs' weight relative to RCTs impact the pooled estimates (log odds ratios). All statistical approaches were implemented with Bayesian random-effects meta-regression models. To compare the statistical approaches, we visualized their pooled estimates in a forest plot with log odds ratios (logORs) and 95% credible intervals (CrIs), and visualized the impact of changed assumptions on how to downweigh NRSs using line charts.

## Results

The pooled estimates obtained using the three approaches were consistent in CS1 and CS4, where study numbers (14; 23) and RCT proportions (57%, 87%) were relatively large, but were less consistent otherwise. However, in all the four cases, the hierarchical modelling approach resulted in much larger 95% confidence intervals than the other two approaches. Covariates did not impact the pooled estimates significantly, but for the hierarchical modelling approach, the pooled estimates varied greatly with the changed NRSs' weight.

## Conclusions

The power prior approach is more reliable than the naive pooling and hierarchical modelling approach, for synthesizing evidence from RCTs and NRSs in a meta-analysis, but none of the approaches could guarantee the accuracy of pooled estimates when the number or proportion of RCTs is small. Further research is needed to confirm our findings in more cases across different disease fields, and to identify scenarios where the hierarchical modelling approach could be more reliable.

## Introduction

Meta-analysis is a research method to estimate treatment effects of healthcare interventions by systematically synthesizing findings from single studies (1,2). Given the strengths of meta-analyses over single studies, such as increased statistical power, meta-analyses which use randomized controlled trials (RCTs) as data sources are often placed at the top of an evidence pyramid to inform healthcare decision-making, in clinical, regulatory, or health technology assessment (HTA) settings (3-5). For a meta-analysis, RCTs and non-randomized studies (NRSs) are data sources that may complement each other. RCTs are normally considered as the gold-standard, as randomization reduces bias and rigorously examines cause-effect relationships (6), but face the drawback of strict experimental settings or eligibility criteria for patients (7). Therefore, with combined estimates from RCTs and NRSs, decision-makers may gain more confidence that experimental findings can be extrapolated to real-world populations (8). In addition, the increased evidence from NRSs is especially useful for decision-making on rare diseases or when conducting an RCT is infeasible (9). While combining RCTs and NRSs is promising, validity of such meta-analyses can be questioned, due to quality concerns of included single studies and heterogeneity among them. For example, compared to RCTs, NRSs tend to have higher risk of bias (e.g. confounding) and less strict patient inclusion criteria (10). Inclusion of such NRSs without any adjustment in a meta-analysis would lead to biased estimates and questionable conclusions. Consequently, NRSs are often considered as complementary evidence, and there is no priori interest in pooling NRSs with RCTs.



To address validity concerns of meta-analyses that combine RCTs and NRSs, some statistical approaches have been developed (11-13). According to Jenkins et al. (9), such approaches may be divided in three categories: naïve pooling, power prior, and hierarchical modelling. The naive pooling approach is the least recommended, and only used as a reference approach to investigate effects of including NRSs (13). The reason is that this approach combines RCTs and NRSs without considering their heterogeneity in design or risk of bias (9). Regarding the power prior approach, NRSs are downweighed by acting as prior information in a Bayesian meta-analysis (BMA), e.g., as prior probability distribution on mean or variance of treatment effects (9).To gain prior information, a meta-analysis using merely NRSs should be first conducted (13). In contrast, the hierarchical model approach includes RCTs and NRSs in one BMA, but downweighs NRSs through a BMA model with multiple hierarchical levels, which considers heterogeneity of treatment effects within and across study designs (9).

While these approaches supporting synthesis of RCTs and NRSs have emerged, they were seldom applied or compared in case studies, partly due to lack of a tool that can be used for the comparison. The approaches were developed by statisticians in various studies, and were often separately available in different softwares (e.g. WinBUGS, STATA, and R) (14), which made the comparison technically difficult. The Crossnma package (Version 1.0.1), a recently developed R package by Hamza et al. (15) within the HTx project (16), may partly overcome this difficulty, as it enables comparison of the three categories of approaches in one software. Also, it supports investigating impact of assumptions related to the approaches (e.g. how to downweigh NRSs). However, since the Crossnma package was developed recently, it has not yet been applied in case studies.

Hence, the aim of this study was to apply the Crossnma package in case studies, and to compare whether the naïve pooling, power prior, and hierarchical modelling approaches provided consistent pooled estimates and credible intervals. We chose cases in the field of myelodysplastic syndromes (MDS) and diabetes. MDS is a group of cancers in which blood-forming cells in the bone marrow do not mature or function properly (17).Since MDS is a relatively rare disease, with an incidence of four cases per 100 000 individuals per year (18,19), including all available evidence, such as NRSs, in a meta-analysis could improve the probability to achieve sufficient power to detect a certain true effect and to draw a credible conclusion (20). Also, since diabetes mellitus is a chronic disease with a large number of healthcare interventions, long-term data (e.g. beyond 5 years) and evidence that compares a pair of intervention is often lacking (21-23).The data obtained from NRSs, which are relatively cost-efficient, could complement RCTs, especially when RCTs are less feasible to conduct (24,25). This research was performed as part of the HTx project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825162 (16).

# Methods

## Case preparation & description

We applied statistical approaches used for synthesizing RCTs and NRSs in a meta-analysis to two cases of MDS (26-28) and two cases of diabetes (29,30). The target population, intervention, comparator, outcome, study characteristics, and references of each case are shown in Table 1. The target intervention or comparator in Case study 1 (CS1) was a type of allogeneic hematopoietic stem cell transplantation (allo-HSCT), a medical procedure to treat MDS patients. The second case study (CS2) also focused on

MDS, and the target intervention was the Iron chelation therapy (ICT), a non-invasive method for evaluating the degree of iron overload, a common clinical problem (31). In Case study 3 (CS3) and Case study 4 (CS4), anti-diabetic medications, i.e., metformin and glucagon-like peptide-1 (GLP-1) receptor agonists, were considered as the target interventions, and the incidences of pancreatic cancer and heart failure were target outcomes, respectively. According to publications related to the cases, the proportion of RCTs differed among the four cases, ranging from 14% to 87%.

Quality judgement of RCTs had already been made published in all the four cases. In CS1,CS2, and CS4, RCTs had been graded as "low" or "high", using the Cochrane's risk of bias tool (Cochrane) (32) , while in CS3, RCTs had been graded from "Point one" to "Point five" (a larger point indicating better methodological quality), using the Jadad scale (33). Regarding NRSs, CS1 did not provide relevant quality judgement. In CS2 and CS3, NRSs had been graded as "good", "fair", or "poor", using the Newcastle – Ottawa scale (NOS) (34), while in CS4, NRSs had been graded as "high", "moderate", "low", or "very low", using the checklist derived from both the NOS and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) checklist (35).

For the convenience of describing the quality of RCTs or NRSs, obtained from various case studies, we labeled the quality of an RCT as "high", if it was graded as "high", using the Cochrane, or as "Five Point", using the Jadad scale. Also, we labeled the quality of an NRS as "high", if it was graded as "good" using the NOS, or as "high" in CS4. The quality of NRSs in CS1, which had not been judged, was labeled as "unclear". All the other primary studies were labelled as having "low" quality.

Regarding data collection, most data had been reported by case authors, while data that were not reported, i.e., the number of events or no-events in CS2 and CS4, were collected by one researcher (LJ) from the primary studies. Any missing data on covariates were imputed with medians. The raw data for all the four cases are available upon request.

**Table 1. Description of the four cases**

| | Case study 1 | Case study 2 | Case study 3 | Case study 4 |
|---|---|---|---|---|
| Target population | Patients with acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS) | Patients with low-to-intermediate-risk myelodysplastic syndrome (MDS) | Patients with type 1 or type 2 diabetes mellitus | Adult patients with type 2 diabetes |
| Intervention | Reduced-intensity conditioning (RIC) allogeneic hematopoietic stem cell transplantation (alloHSCT) | Iron chelation therapy (ICT) | Metformin | Glucagon-like peptide-1 (GLP-1) receptor agonists |
| Comparator | Myeloablative conditioning (MAC) alloHSCT | Non-user | Non-user | Non-user |
| Outcome | Overall survival at the end-of-follow-up | Overall mortality | Incidence of pancreatic cancer (PaC) | Incidence of heart failure |
| Number of RCTs/ NRSs | 8 / 6 | 1 / 7 | 2 / 4 | 20 / 3 |
| Tool to appraise quality of RCTs/ NRSs | Cochrane / None | Cochrane / NOS | Jadad scale / NOS | Cochrane / NOS + GRADE |
| Reference | Song 2021 (26) / Zeng 2014 (27) | Yang 2021 (28) | Singh 2013 (29) | Li 2016 (30) |

## Sensitivity analyses

An overview of sensitivity analyses that were applied across the four cases is shown in Appendix 1. In the base case, all statistical approaches were implemented using both RCTs and NRSs as data sources, without covariate adjustment. To investigate whether a pooled estimate using only RCTs or NRSs differed significantly from the estimates using both two sources, we repeated the naïve pooling approaches using only RCTs or NRSs. Also, as study covariates might impact pooled estimates, we repeated all identified statistical approaches with adjustment of age and duration of follow-up for comparison.

Regarding the power prior and hierarchical modelling approach, the assumptions on how to downweigh NRSs might impact pooled estimates. According to the Crossnma package, the power prior approach downweighs NRSs by inflating their variances of

prior distribution, while the hierarchical modelling approach downweighs NRSs, by analyzing RCTs and NRSs separately before merging them with different weights (15). To investigate the impact of such assumptions on pooled logORs, we adjusted assumed parameter value in sensitivity analyses. More specifically, for both approaches, we adjusted the weight of NRSs relative to RCTs, from 20% to 90% (60% in the base case). For the hierarchical modelling approach, only CS2 and CS3 were included in the sensitivity analysis. The reason was that the hierarchical modelling approach provided by the Crossnma package, was only available, when RoB of all studies was clearly reported and studies with both high and low RoB existed. CS1 included a large proportion of studies with unclear RoB, while CS4 only included high-RoB studies.

### Statistical analysis

If the log OR estimator is not defined, i.e., with at least one cell with a frequency of zero (a "zero cell") in the corresponding 2 × 2 table (36), we corrected for the "zero cell", by adding 1 to each cell in the 2 × 2 table. All statistical approaches were implemented with Bayesian random-effects meta-regression models. As a starting point, we ran Markov chain Monte Carlo (MCMC) simulations with 4 chains, 50000 samples, 20000 burn-ins, and without thinning (i.e. thin = 1). The MCMC convergence was tested using the Gelman-Rubin statistic (Rhat), and was considered acceptable if Rhat was less than 1.1 (37). If the chains did not converge, we reran the simulation by increasing samples, burn-ins, or thinned out the chains, until they converged. We ran all statistical approaches using the R software (Version 3.4.2), with the Crossnma package (38).The package (version 1.0.1) was published in April 2022 on the Comprehensive R Archive Network (CRAN) website (39).To compare the statistical approaches, we visualized their pooled estimates in a forest plot with log odds ratios (logORs) and 95% credible intervals (CrIs). In addition, we visualized the impact of changed assumptions of the power prior and hierarchical modelling approach on how to downweigh NRSs using line charts.



## Results

### Main analysis

Pooled estimates and credible intervals (Crls) obtained using the three statistical approaches (i.e. naïve pooling, power prior, and hierarchical modelling) in the two MDS cases (i.e. CS1 and CS2) are shown in Figure 1.

In CS1, the pooled logOR obtained using RCTs as data source was the same as that using NRSs (-0.06). Also, all the three statistical approaches provided similar pooled logOR (-0.08, -0.07, and -0.08). The 95% credible intervals were almost the same

among the three approaches (ranging between -0.32 and 0.16). According to results obtained with all of the three approaches, reduced-intensity conditioning alloHSCT, as compared to myeloablative conditioning alloHSCT, was not associated with an increased or decreased risk of mortality. After covariate adjustment, the pooled logORs using the naïve pooling and hierarchical modelling approach became larger, but with larger 95% CrIs.

In CS2, although the pooled logOR obtained using NRSs as data sources were smaller than that from RCTs, no significant difference in pooled estimates was detected, due to the smaller number of RCTs and thus large uncertainty (95% CrI = -4.53, 4.31). Also, the logORs obtained using the power prior and naïve pooling approach were similar (-0.49 and -0.57), while the uncertainty relevant to the power prior approach was relatively smaller (95% CrI = -0.8, -0.17). According to the two approaches, the iron chelation therapy, as compared to non-users, was associated with lower incidence of mortality. In contrast, the pooled logOR obtained using the hierarchical modelling approach (-0.03) was larger than those with the other two approaches, and it involved so great uncertainty that no conclusion could be drawn. After adjusting for covariates, the pooled estimates became larger, using the naïve pooling and hierarchical modelling approach, while the pooled estimates did not vary, using the power prior approach.

The pooled estimates and relevant uncertainty on two cases of diabetes (CS3 and CS4) are shown in Figure 2. In CS3, due to the large 95% CrIs, the pooled logORs obtained using either RCTs or NRSs as data sources could not be compared. Also, while the pooled logORs obtained using the three approaches were similar, ranging between -0.28 and -0.14, the 95% CrIs relevant to the power prior approach was much smaller, compared to the other two approaches. According to the power prior approach, metformin was associated with a slightly lower risk of pancreatic cancer. In contrast, according to the other two approaches, no conclusive results could be obtained. Covariate adjustment did not alter the interpretation of results.

In CS4, due to the large 95% CrIs, the pooled logORs obtained using either RCTs or NRSs as data sources could not be compared. All the three approaches provided similar logORs (-0.49, -0.52, and -0.5) and relevant uncertainty (all the 95% CrIs ranging approximately between -1 and 0). Accordingly, glucagon-like peptide-1 receptor agonists were associated with low risk of heart failure. After covariate adjustment, interpretation on efficacy of GLP-1 was affected, with larger 95% CrIs, indicating some chance that GLP-1 did not lower the risk of heart failure. Still, in CS4, covariate adjustment had similar impact on results and interpretation, obtained using all the three approaches.

| Case study 1 (8 RCTs; 6 NRSs) | Sample | Log Odds Ratio | Lower limit | Upper limit | Log Odds Ratio (95% CI) |
|---|---|---|---|---|---|
| Shimoni 2012 | 191 | -0.25 | -0.74 | 0.24 | |
| Philipp 2010 | 93 | 0.28 | -0.49 | 1.06 | |
| Martino 2012 | 844 | -0.11 | -0.45 | 0.22 | |
| Luger 2012 | 4772 | -0.04 | -0.15 | 0.07 | |
| Khabori 2011 | 101 | 0.12 | -0.58 | 0.81 | |
| Hiramoto 2014 | 115 | -0.21 | -0.88 | 0.47 | |
| Lioure 2012 | 164 | -0.01 | -0.70 | 0.68 | |
| Bornhauser 2012 | 184 | 0.12 | -0.42 | 0.66 | |
| Scott 2017 | 272 | -0.50 | -1.03 | 0.02 | |
| Ringden 2013 | 37 | 0.71 | -0.67 | 2.10 | |
| MC-FludT.14/L.Trial I 2018 | 320 | 0.19 | -0.23 | 0.61 | |
| Kroger 2017 | 129 | 0.69 | -0.04 | 1.42 | |
| Bornhauser 2012 | 195 | -0.23 | -0.68 | 0.22 | |
| Beelen 2019 | 460 | -0.66 | -1.02 | -0.30 | |
| ***Overall*** | | | | | |
| Naïve pooling - RCTs only as data source | | -0.06 | -0.50 | 0.38 | |
| Naïve pooling - NRSs only as data source | | -0.06 | -0.31 | 0.19 | |
| Naïve pooling | | -0.08 | -0.28 | 0.12 | |
| Power prior | | -0.07 | -0.30 | 0.16 | |
| Hierarchical modelling | | -0.08 | -0.32 | 0.16 | |
| Naïve pooling - adjusted by age and duration of follow-up | | 0.34 | -0.17 | 0.86 | |
| Power prior - adjusted by age and duration of follow-up | | 0.02 | -0.26 | 0.31 | |
| Hierarchical modelling - adjusted by age and duration of follow-up | | 0.34 | -0.20 | 0.88 | |

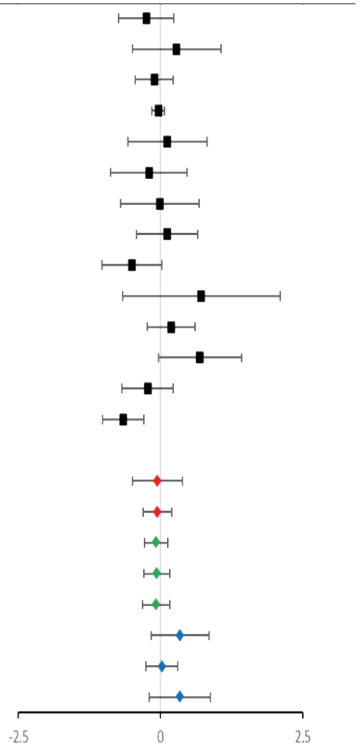



**Figure 1. Forest plots to estimate the outcomes of interest in patients with myelodysplastic syndrome (Case study 1 & 2).**

Red color indicates the naive pooling approach was applied using one data source only, i.e., either randomized controlled trials or non-randomized studies; Green, a statistical approach was applied using both data sources; Blue, a statistical approach was applied using both data sources, with adjustment of two covariates, i.e., age and duration of follow-up.

| Case study 2 (1 RCT; 6 NRSs) | Sample | Log Odds Ratio | Lower limit | Upper limit | Log Odds Ratio (95% CI) |
|---|---|---|---|---|---|
| Angelucci 2020 | 225 | -0.21 | -0.67 | 0.24 | |
| Rose 2010 | 97 | -2.02 | -3.05 | -0.98 | |
| Lyons 2017 | 599 | -0.51 | -0.84 | -0.18 | |
| Delforge 2014 | 127 | -1.26 | -1.96 | -0.57 | |
| Wong 2018 | 182 | -0.05 | -0.53 | 0.43 | |
| Raptis 2010 | 78 | 0.28 | -0.42 | 0.99 | |
| Leitch 2012 | 182 | -0.41 | -0.97 | 0.16 | |
| Overall | | | | | |
| Naïve pooling - RCTs only as data source | | -0.11 | -4.53 | 4.31 | |
| Naïve pooling - NRSs only as data source | | -0.65 | -1.60 | 0.30 | |
| Naïve pooling | | -0.57 | -1.29 | 0.14 | |
| Power prior | | -0.49 | -0.80 | -0.17 | |
| Hierarchical modelling | | -0.03 | -24.25 | 24.18 | |
| Naïve pooling - adjusted by age and duration of follow-up | | -0.10 | -1.59 | 1.38 | |
| Power prior - adjusted by age and duration of follow-up | | -0.58 | -0.94 | -0.21 | |
| Hierarchical modelling - adjusted by age and duration of follow-up | | 0.78 | -26.74 | 28.30 | |

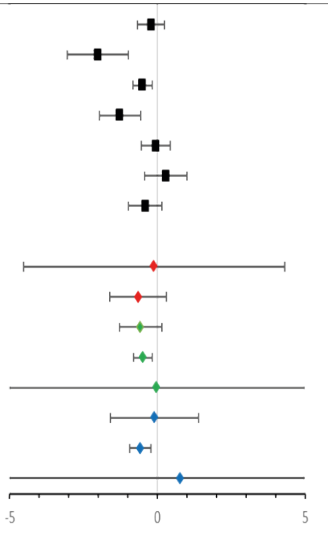



**Figure 1. Continued**

| Case Study 3 (2 RCTs; 4 NRSs) | Sample | Log Odds Ratio | Lower limit | Upper limit | Log Odds Ratio (95% CI) |
|---|---|---|---|---|---|
| Bodmer 2012 | 19341 | 0.33 | 0.22 | 0.44 | |
| Liao 2012 | 49803 | 0.14 | 0.11 | 0.18 | |
| Ruiter et al 2012 | 85289 | -1.05 | -1.07 | -1.03 | |
| Li 2009 | 361 | -1.00 | -1.47 | -0.54 | |
| ADOPT 2006 | 4351 | -0.10 | -0.19 | -0.01 | |
| RECORD 2009 | 2225 | 0.20 | 0.08 | 0.32 | |
| ***Overall*** | | | | | |
| Naïve pooling - RCTs only as data source | | 0.04 | -2.41 | 2.49 | |
| Naïve pooling - NRSs only as data source | | -0.40 | -2.23 | 1.43 | |
| Naïve pooling | | -0.25 | -1.04 | 0.55 | |
| Power prior | | -0.14 | -0.37 | 0.09 | |
| Hierarchical modelling | | -0.28 | -11.66 | 11.11 | |
| Naïve pooling - adjusted by age and duration of follow-up | | -1.26 | -7.14 | 4.61 | |
| Power prior - adjusted by age and duration of follow-up | | -0.37 | -0.72 | -0.03 | |
| Hierarchical modelling - adjusted by age and duration of follow-up | | -1.04 | -14.52 | 12.44 | |

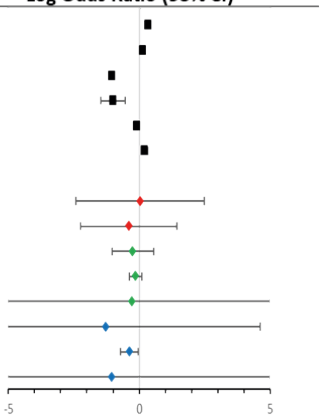



**Figure 2. Forest plot to estimate outcomes in patients with myelodysplastic syndrome (Case study 3 & 4).**
Red color indicates the naive pooling approach was applied using one data source only, i.e., either randomized controlled trials or non-randomized studies; Green, a statistical approach was applied using both data sources; Blue, a statistical approach was applied using both data sources, with adjustment of two covariates, i.e., age and duration of follow-up.

| Case Study 4 (20 RCTs; 3 NRSs) | Sample | Log Odds Ratio | Lower limit | Upper limit | Log Odds Ratio (95% CI) |
|---|---|---|---|---|---|
| Inagaki 2012 | 431 | 0.68 | 0.42 | 0.95 | |
| NCT00294723 2010 | 749 | 0.00 | -0.21 | 0.22 | |
| NCT00318461 2010 | 849 | -1.09 | -1.36 | -0.82 | |
| NCT00360334 2009 | 970 | -0.40 | -0.60 | -0.19 | |
| NCT00614120 2010 | 238 | -0.72 | -1.08 | -0.36 | |
| NCT00701935 2013 | 932 | -0.41 | -0.62 | -0.20 | |
| NCT00838903 2014 | 84 | -0.87 | -1.48 | -0.25 | |
| NCT00838916 2014 | 407 | 0.02 | -0.30 | 0.33 | |
| NCT00839527 2014 | 609 | 0.71 | 0.49 | 0.94 | |
| NCT00849017 2014 | 745 | -0.74 | -0.96 | -0.52 | |
| NCT00849056 2014 | 390 | -1.55 | -1.86 | -1.25 | |
| NCT00855439 2015 | 552 | -1.60 | -1.84 | -1.37 | |
| NCT00960661 2013 | 301 | -1.39 | -1.73 | -1.05 | |
| NCT01064687 2015 | 305 | -0.69 | -1.01 | -0.37 | |
| NCT01075282 2015 | 46 | 0.09 | -0.74 | 0.92 | |
| NCT01126580 2015 | 631 | -0.71 | -0.93 | -0.48 | |
| NCT01191268 2014 | 704 | -0.68 | -0.94 | -0.42 | |
| NCT01512108 2014 | 423 | -0.68 | -0.96 | -0.39 | |
| NCT01620489 2014 | 807 | 0.37 | 0.16 | 0.58 | |
| Pratley 2013 | 811 | 0.00 | -0.21 | 0.20 | |
| NCT01060059 2013 | 364 | 0.01 | -0.30 | 0.32 | |
| Kannan 2015 | 281 | 0.68 | 0.35 | 1.01 | |
| Paul 2015 | 755 | -1.76 | -1.97 | -1.55 | |
| ***Overall*** | | | | | |
| Naïve pooling - RCTs only as data source | | -0.50 | -1.05 | 0.05 | |
| Naïve pooling - NRSs only as data source | | -0.37 | -4.59 | 3.85 | |
| Naïve pooling | | -0.49 | -1.00 | 0.01 | |
| Power prior | | -0.52 | -1.05 | 0.01 | |
| Hierarchical modelling | | -0.50 | -1.01 | 0.00 | |
| Naïve pooling - adjusted by age and duration of follow-up | | -0.69 | -1.74 | 0.35 | |
| Power prior - adjusted by age and duration of follow-up | | -0.60 | -1.58 | 0.37 | |
| Hierarchical modelling - adjusted by age and duration of follow-up | | -0.69 | -1.73 | 0.35 | |

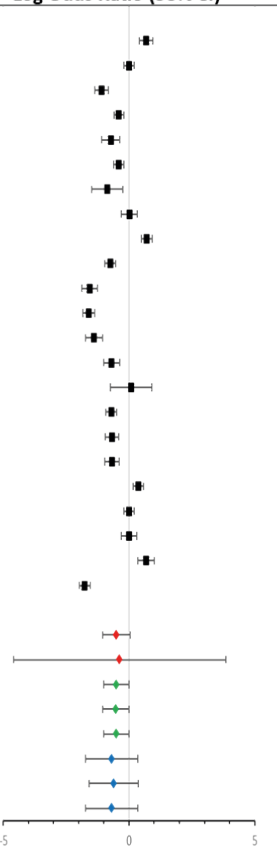



**Figure 2. Continued**

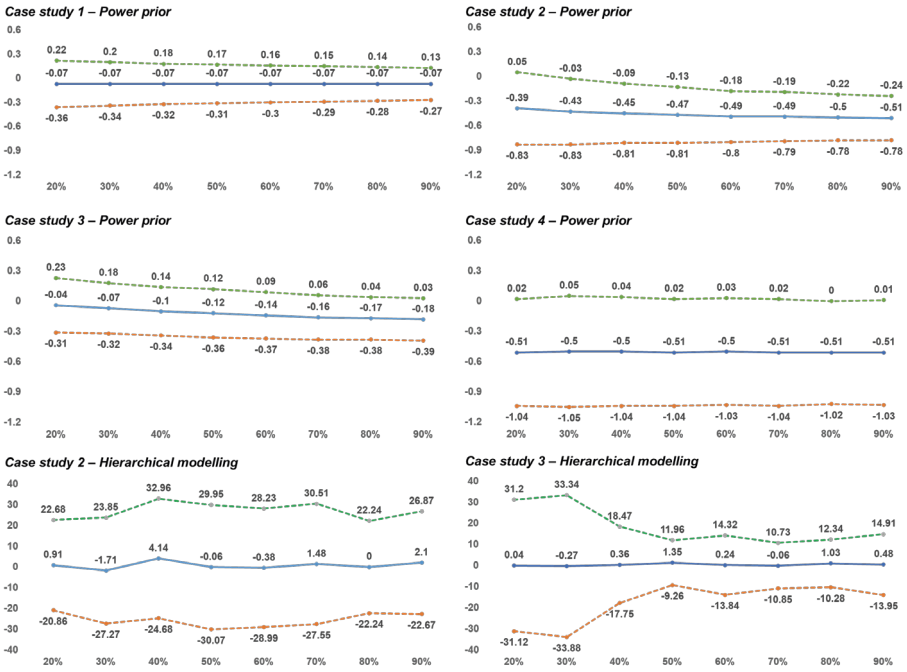



**Figure 3. Sensitivity analysis of assumptions on downweighing NRSs, using the power prior or the hierarchical modelling approach.**
A higher percentage on the X-axis indicates a higher weight of NRSs relative to RCTs: 0, NRSs not taken into account; 100%, NRSs with the same weight as RCTs. The solid blue line indicates the pooled log odds ratio; the dotted green line, the upper 95% credible intervals; the dotted red line, the lower 95% credible intervals

8

In summary, the pooled estimates and 95% CrIs using the naïve pooling, power prior, and hierarchical modelling approach were similar, when the study number was relatively high (CS1 and CS4), but varied otherwise. In particular, the hierarchical modelling approach could provide a large credible interval that nullifies an intervention-efficacy association detected using the power prior approach. In addition, results and interpretations obtained using the three approaches might be influenced by covariate adjustment, but the impact was similar among the three.

## Sensitivity analysis on assumptions to downweigh NRSs

Figure 3 illustrates the impact of assumptions of how to downweigh NRSs on pooled logORs, using either the power prior or the hierarchical modelling approach. Regarding the power prior approach, impact of the NRSs' weight on results depended on case studies. In CS1 and CS4, the pooled logORs were not sensitive to the assumption, but in CS2 and CS3, the pooled logORs decreased slightly, as the NRSs' weight increased. In

all the four cases, the 95% credible intervals narrowed down slightly, indicating smaller uncertainty, as NRSs were assigned a larger weight. While the pooled logORs and relevant uncertainty might change, the slight change did not alter the interpretation of results in the basic scenario. Regarding the hierarchical modelling approach, the pooled logORs varied significantly, with extremely large 95% Crls. Because of the great uncertainty related to pooled estimates, we could not affirm or deny any association between the pooled estimates and the changed weight.

## Discussion

We implemented three statistical approaches which were designed for synthesizing randomized controlled trials and non-randomized studies in a meta-analysis (i.e. the naïve pooling, power prior, and hierarchical modelling approach), in two cases of myelodysplastic syndrome and diabetes. We compared the pooled estimates and relevant uncertainty, and explored how the assumptions on downweighing NRSs could impact the results. Our study showed that the results obtained using the three approaches were consistent, when the number of studies was relatively large, but were less consistent otherwise. In particular, the hierarchical modelling approach resulted in much larger uncertainty than the power prior approach. In addition, our study showed that, the assumption on NRSs' weight for the power prior approach did not have a large impact on pooled estimates, but we could not judge the assumption for the hierarchical modelling approach, due to large uncertainty on pooled estimates.

Our results were almost consistent with results from Yao et al. 2023 (40).They wrote their own codes for the naïve pooling, power prior, and hierarchical approaches with a statistical modeling platform (i.e. Stan), and implemented these approaches in two cases of rare events meta-analysis, using the RStan package, the R interface to Stan (41). Also, they compared the pooled estimates and credible intervals, and investigated the impact of the approaches' assumptions on how NRSs were downweighed (i.e., inflated variance of prior distribution in their cases) on pooled estimates. They found that results from different approaches were inconsistent and that the three-level hierarchical modelling approach they applied had much larger uncertainty (credible intervals) than the other approaches. Their sensitivity analysis also showed that the changed weight of NRSs could impact pooled estimates, using either the power prior and hierarchical modelling approach. In our research, we could not detect the association between the NRS weight and pooled estimates, for the hierarchical modelling approach. One possible explanation is the smaller number of studies

included in our CS2 and CS3 (n=8;6), as compared to the two cases from the study by Yao et al. (n=11;16).

We found that the Crossnma package had potential to be applied by users without expertise in statistics. Originally, Hamza et al. 2023 developed the Crossnma package, and proved the feasibility of comparing the statistical approaches in an R package (15). In our study, the manual of the Crossnma package was strictly followed, and assumptions on how to downweigh NRSs were explored by altering the package-specific R arguments (e.g. "down.wgt" to downweigh high-ROB studies using the hierarchical modelling approach). Also, the R arguments used in the Crossnma package were mainly basic concepts related to a meta-analysis, such as meta-regression and covariates, while the package avoided the use of expressions that might only be understood with statistical backgrounds (38,39). With such features, the implementation and comparison of these statistical approaches, as well as interpretation of results, with the Crossnma package, did not necessarily demand a strong knowledge background in statistics.

The increased user-friendliness of approaches to synthesizing RCTs and NRSs might lay the roadmap for promoting real-world evidence in decision-making. Regarding regulatory decision-making, NRSs have already been used or could potentially be used in some scenarios (e.g. to support a supplemental indication or adaptive approval of drugs) (42). Still, novel approaches are needed, so regulators could balance the advantages (e.g. powerful analyses for rare diseases) and disadvantages (e.g. lack of controlled measurements) of NRSs (42). Similarly, conducting meta-analyses is often a step for establishing evidence for HTA (43), but consensus is lacking on how to conduct a meta-analysis incorporating real-world evidence (44). For HTA stakeholders, an approach that synthesizes RCTs and NRSs, complemented by a manual that takes the user-friendliness into account, might be a solution, and could be promoted in future HTA guidelines.



Another implication of our study was that implementation of statistical approaches that synthesize RCTs and NRSs could enhance complementation of the two data sources, by overcoming some of their limitations. In the original article of CS3, only NRSs were included in the final meta-analysis, as the authors considered the two identified RCTs heterogeneous, e.g. in terms of duration of follow-up (29). In our study, by synthesizing RCTs and NRSs while assuming their heterogeneity using the hierarchical modelling approach, we confirmed findings from CS3 that the association of metformin and risk of pancreatic cancer could not be detected, due to considerable heterogeneity among the studies. Similarly, CS1 was made up of two separate meta-analyses. Song

et al (26) stated that one limitation of their meta-analysis was the small number of studies, as they only included six RCTs. In contrast, Zeng et al (27) considered the high proportion of NRSs in their meta-analysis a limitation. By merging all primary studies and downweighing NRSs, we doubled the number of included studies, and reduced the proportion of NRSs. According to previous research, consensus has been made that RCTs and NRSs are complementary in decision-making, but consensus is still lacking on methods to enhance the complementation (e.g. methods to generate or synthesize evidence) and their relevant impact (45). The power prior approach has potential and its impact on findings and limitations of meta-analyses need to be investigated in future research. Moreover, we recommend stakeholders to combine RCTs and NRSs in a meta-analysis, when the total number of studies is relatively large (e.g. more than 10), and when RCTs are more than NRSs. Otherwise, it would be reasonable to synthesize NRSs separately, and to consider NRSs only as a complementary source of evidence for the decision-making purpose.

## Limitations

Our study has several limitations. One limitation was that we did not address the "zero-cell" problem that occurs in a rare-events meta-analysis, by following the common practice (i.e. adding 0.5 to the numbers of event and no-event of primary studies involving this problem) (36). The reason was that the Crossnma package (Version 1.0.1) only accepted integers. Another limitation was that we did not include a hierarchical modelling approach such as developed by Verde et al (46), which features a model term to adjust for internal validity bias. Though it was included in the Crossnma package, we failed to run this approach, as system errors sometimes occurred during minor updates of the Crossnma package (Version 1.0.1). Still, the user-friendliness brought by this package makes it a promising tool for application in the future. Therefore, we hope to see the continuous improvement of Crossnma package, in terms of approach reliability and, if available, inclusion of additional statistical approaches. Finally, the number of cases included in our study was quite small (n=4), and only binary outcomes, in format of odds ratio, were investigated. To test transferability of our findings, we recommend future research comparing the statistical approaches in more cases.

## Conclusions

The power prior approach is more reliable than the naïve pooling and hierarchical modelling approach, for synthesizing evidence from RCTs and NRSs in a meta-analyses. Further research is needed to confirm our findings in more cases across different disease fields, and to investigate the scenarios where the power prior approach could be more reliable.

## Author contribution

LJ collected data, conduct statistical analysis, and wrote the draft manuscript. All co-authors contributed to this study and critically reviewed and approved the manuscript.

# References


1. Shorten A, Shorten B. What is meta-analysis?. Evid Based Nurs. 2013 Jan 1;16(1):3-4.
2. Haidich AB. Meta-analysis in medical research. Hippokratia. 2010 Dec;14(Suppl 1):29.
3. Velasco-Garrido M, Busse R. Assessing research. In: Health technology assessment: an introduction to objectives, role of evidence, and structure in Europe. World Health Organization. 2005.
4. Schlegl E, Ducournau P, Ruof J. Different weights of the evidence-based medicine triad in regulatory, health technology assessment, and clinical decision making. Pharmaceut Med. 2017;31(4):213-6.
5. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. BMJ Evid Based Med. 2016 Aug 1;21(4):125-7.
6. Hariton E, Locascio JJ. Randomised controlled trials—the gold standard for effectiveness research. BJOG. 2018 Dec;125(13):1716.
7. de Vera-Gómez PR, Mateo-Rodríguez C, Vivas-López C, Serrano-Olmedo I, Méndez-Muros M, Morales-Portillo C, Jiménez MS, Hernández-Herrero C, Martínez-Brocca MA. Effectiveness of a flash glucose monitoring systems implementation program through a group and telematic educational intervention in adults with type 1 diabetes. Endocrinología, Diabetes y Nutrición (English ed.). 2022 Nov 1;69(9):657-68.
8. Briere JB, Bowrin K, Taieb V, Millier A, Toumi M, Coleman C. Meta-analyses using real-world data to generate clinical and epidemiological evidence: a systematic literature review of existing recommendations. Curr Med Res Opin. 2018 Dec 2;34(12):2125-30.
9. Jenkins DA, Hussein H, Martina R, Dequen-O'Byrne P, Abrams KR, Bujkiewicz S. Methods for the inclusion of real-world evidence in network meta-analysis. BMC Med Res Methodol. 2021 Dec;21(1):1-9.
10. Metelli S, Chaimani A. Challenges in meta-analyses with observational studies. BMJ Ment Health. 2020 May 1;23(2):83-7.
11. Hamza T, Chalkou K, Pellegrini F, et al. Synthesizing cross-design evidence and cross-format data using network meta-regression. Res Synth Methods. 2023 Mar;14(2):283-300..
12. Schmitz S, Adams R, Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. Stat Med. 2013 Jul 30;32(17):2935-49.
13. Efthimiou O, Mavridis D, Debray TP, et al. Combining randomized and non-randomized evidence in network meta-analysis. Stat Med. 2017 Apr 15;36(8):1210-26.
14. Zhang K, Arora P, Sati N, et al. Characteristics and methods of incorporating randomized and nonrandomized evidence in network meta-analyses: a scoping review. J Clin Epidemiol. 2019 Sep 1;113:1-0.
15. Hamza T, Chalkou K, Pellegrini F, et al. Synthesizing cross-design evidence and cross-format data using network meta-regression. Res Synth Methods. 2023 Jan 10.
16. HTx: About HTx project. Available from: https://www.htx-h2020.eu/about-htx-project. [Accessed Oct 25, 2022]
17. Adès L, Itzykson R, Fenaux P. Myelodysplastic syndromes. Lancet. 2014 Jun 28;383(9936):2239-52.
18. Fenaux P, Haase D, Sanz GF, Santini V, Buske C. Myelodysplastic syndromes: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology. 2014 Sep 1;25:iii57-69.
19. University of Groningen. Real-world studies of patients with myelodysplastic syndromes: survival, treatment and outcomes in a population-based setting. Available from: https://research.rug.nl/en/publications/real-world-studies-of-patients-with-myelodysplastic-syndromes-sur. [Accessed Oct 27, 2023]

20. Jia P, Lin L, Kwong JS, Xu C. Many meta-analyses of rare events in the Cochrane Database of Systematic Reviews were underpowered. J Clin Epidemiol. 2021 Mar 1;131:113-22.

21. Centers for Disease Control and Prevention (CDC). What is Diabetes? Available from : https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=Diabetes%20is%20a%20chronic%20(long,your%20pancreas%20to%20release%20insulin. [Accessed 16 Apr, 2023]

22. Nathan DM. Diabetes: advances in diagnosis and treatment. Jama. 2015 Sep 8;314(10):1052-62.

23. Kamusheva M, Tachkov K, Dimitrova M, et al. A systematic review of collective evidences investigating the effect of diabetes monitoring systems and their application in health care. Front Endocrinol (Lausanne). 2021 Mar 16;12:636959.

24. Zaccardi F, Davies MJ, Khunti K. The present and future scope of real-world evidence research in diabetes: what questions can and cannot be answered and what might be possible in the future?. Diabetes Obes Metab. 2020 Apr;22:21-34.

25. Burcu M, Manzano-Salgado CB, Butler AM, Christian JB. A Framework for Extension Studies Using Real-World Data to Examine Long-Term Safety and Effectiveness. Ther Innov Regul Sci. 2022 Jan;56(1):15-22.

26. Song Y, Yin Z, Ding J, Wu T. Reduced intensity conditioning followed by allogeneic hematopoietic stem cell transplantation is a good choice for Acute myeloid leukemia and myelodysplastic syndrome: a Meta-analysis of Randomized controlled trials. Front Oncol. 2021 Oct 7;11:708727.

27. Zeng W, Huang L, Meng F, Liu Z, Zhou J, Sun H. Reduced-intensity and myeloablative conditioning allogeneic hematopoietic stem cell transplantation in patients with acute myeloid leukemia and myelodysplastic syndrome: a meta-analysis and systematic review. Int J Clin Exp Med. 2014;7(11):4357.

28. Yang S, Zhang MC, Leong R, Mbuagbaw L, Crowther M, Li A. Iron chelation therapy in patients with low-to intermediate-risk myelodysplastic syndrome: A systematic review and meta-analysis. Br J Haematol. 2022 Apr;197(1):e9-11.

29. Singh S, Singh PP, Singh AG, Murad HM, McWilliams RR, Chari ST. Anti-diabetic medications and risk of pancreatic cancer in patients with diabetes mellitus: a systematic review and meta-analysis. Clin Transl Gastroenterol. 2013 Apr 1;108(4):510-9.

30. Li L, Li S, Liu J, et al. Glucagon-like peptide-1 receptor agonists and heart failure in type 2 diabetes: systematic review and meta-analysis of randomized and observational studies. BMC Cardiovasc Disord. 2016 Dec;16:1-4.

31. Poggiali E, Cassinerio E, Zanaboni L, Cappellini MD. An update on iron chelation therapy. Blood Transfus. 2012 Oct;10(4):411

32. Higgins JP, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials.BMJ. 2011 Oct 18;343.

33. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary?. Control Clin Trials. 1996 Feb 1;17(1):1-2.

34. Peterson J, Welch V, Losos M, Tugwell PJ. The Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa: Ottawa Hospital Research Institute. 2011;2(1):1-2.

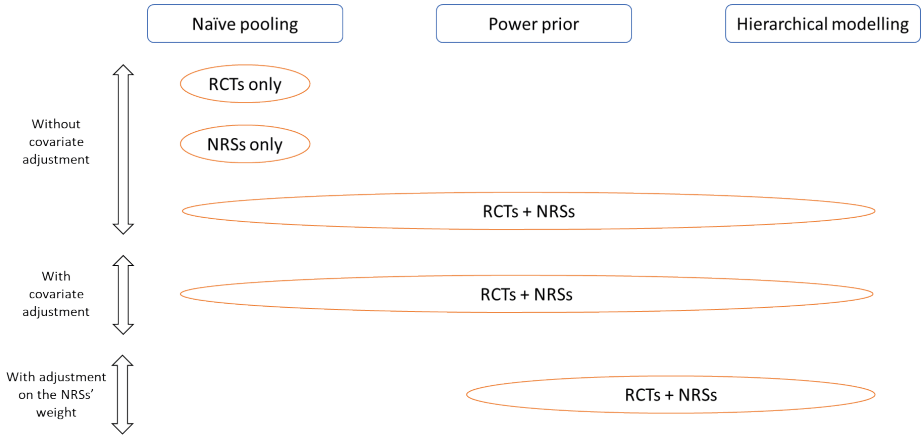
35. BMJ Best Practice. What is GRADE? Available from: https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade. [Accessed 16 April 2023]

36. Weber F, Knapp G, Ickstadt K, Kundt G, Glass Ä. Zero-cell corrections in random-effects meta-analyses. Res Synth Methods. 2020 Nov;11(6):913-9.

37. Vats D, Knudson C. Revisiting the gelman–rubin diagnostic. Stat Sci. 2021 Nov;36(4):518-29.

38. Github. Crosnma. Available from: https://github.com/htx-r/crossnma. [Accessed April 17, 2023]
39. crossnma: Cross-Design & Cross-Format Network Meta-Analysis and Regression . Available from https://cran.r-project.org/web/packages/crossnma/index.html. [Accessed April 17, 2023]
40. Yao M, Wang Y, Mei F, Zou K, Li L, Sun X. Methods for the Inclusion of Real-World Evidence in a Rare Events Meta-Analysis of Randomized Controlled Trials. J Clin Med. 2023 Feb 20;12(4):1690.
41. RStan. Home. Available from: https://mc-stan.org/rstan. [Accessed 26 Nov, 2023]
42. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evauating the use of nonrandomized real-world data analyses for regulatory decision making. Clin Pharmacol Ther. 2019 Apr;105(4):867-77.
43. Chaiyakunapruk N, Saokaew S, Sruamsiri R, Dilokthornsakul P. Systematic review and network meta-analysis in health technology assessment. J Med Assoc Thai. 2014 May 1;97:S33-42.
44. Bowrin K, Briere JB, Levy P, Toumi M, Millier A. Use of real-world evidence in meta-analyses and cost-effectiveness models. J Med Econ. 2020 Oct 2;23(10):1053-60.
45. Morales DR, Arlett P. RCTs and real world evidence are complementary, not alternatives. BMJ. 2023 Apr 3;381.
46. Verde PE. A bias-corrected meta-analysis model for combining, studies of different types and quality. Biom J. 2021 Feb;63(2):406-22.

# Appendices




**Appendix 1. An overview of sensitivity analyses that were applied across the four cases**

Chapter 9

# General Discussion

Given the emergence of novel types of health technologies and the variability of the context of health technology assessment (HTA) (1-4), HTA methods may need to be continuously improved and subsequently used in HTA practice. However, as HTA methods have been developed and implemented repeatedly since 1980s and appeared in large numbers (5,6), some general problems that might affect method development or implementation have occurred over time. As mentioned in the introduction of thesis, the major problems related to HTA method development include the lack of a clear overview of the needs of HTA stakeholders and the limitation of resources, such as available time, high-quality data, and knowledge across research disciplines, while the problems related to method implementation include the lack of expertise or skills to implement a method in the HTA context.

In this thesis, we aimed to provide a framework for how HTA methods should be developed and implemented, by conducting relevant conceptual research, and by illustrating how this framework may improve the process of innovating HTA methods, using the development of HTA methods related to real-world data (RWD) as the cases. In this final chapter, the main results and findings of the previous chapters will be summarized, and in particular relevant implications to the application of real-world data to HTA will be discussed. Also, we discuss solutions to the general problems related to HTA method development and implementation and consider the study limitations and future research opportunities that could further facilitate the development and implementation of new HTA methods.

## Findings & Implications

### Conceptual research on innovation of HTA methods

In the first part of this thesis, we conducted conceptual research to facilitate the understanding of how to innovate HTA methods. In the second chapter, we developed a framework, called the Innovation of HTA Methods (IHTAM) framework, which defines the process of innovating HTA methods and the roles HTA stakeholders could play during the process, through two scoping reviews; iterative brainstorming sessions and discussions among HTA stakeholders within the HTx project. In the third chapter, we explored applicability of the IHTAM framework in three cases of development of novel quantitative methods. According to feedback obtained from the study leaders of those three cases and HTx consortium members, a roadmap was developed to complement the original conceptual framework by addressing some of its limitations, e.g. lack of a checklist to be followed. As mentioned by the case study leaders and HTx consortium

members, the IHTAM framework provided a structural way of thinking, and it was highly relevant to the innovation process of case studies.

The IHTAM framework as developed and used in the second and third chapter includes a generic innovation process consisting of three phases ("Identification", "Development", "Implementation") and nine subphases. In the framework, three roles that HTA stakeholders can play in innovation ("Developers", "Practitioners", "Beneficiaries") are defined. "Developers" indicate stakeholders who develop HTA methods; "Practitioners" indicate stakeholders who implement and use HTA methods; "Beneficiaries" indicate stakeholders who benefit from or are affected by HTA methods. This definition of stakeholders is distinct from the classical definition of HTA stakeholders which mostly focuses on stakeholder groups such as patients, clinicians/healthcare providers, payers, academics and health technology developers.

The IHTAM framework may add value to HTA's good practice, as it could function as a foundation for constructing or improving more specific guidance on innovation, and it could serve as a starting point to illustrate the complex HTA innovation process and how it is related to HTA stakeholders. Moreover, it promotes consideration of key challenges that may exist in innovating HTA methods. Along with the framework, we also provided suggestions on how to use the framework. We highlighted that the steps within the IHTAM framework did not necessarily occur sequentially, and a specific innovation process should always be defined for each method that is innovated.



According to the experienced value of the IHTAM framework, we reasonably speculate that the framework could help reduce the complexity of interdisciplinary collaboration, a general problem of HTA method innovation, in two aspects. From one aspect, the framework emphasizes the importance of identifying the needs for method innovation and evaluating the extent to which the needs are satisfied throughout the innovation process. The increased awareness on the method needs is important, as it could avoid scenarios where method developers only develop a method that may be interesting from an academic perspective but not address any particular needs of the HTA stakeholders. The increased awareness could also help method developers take transferability issues into account from the beginning of method innovation. As the needs are repeatedly emphasized, method developers could be motivated to build closer ties with the previously mentioned practitioners and beneficiaries, who may elaborate on the existing needs.

Another aspect that the IHTAM framework could contribute to in terms of reducing the complexity of interdisciplinary collaboration is to make all classical HTA stakeholders

(e.g. patients, healthcare providers, academics) realize their own potential in method innovation, thus increasing their capability and willingness to participate in the innovation. A good example is the development of patient-reported outcome measures (PROMs), which are for instance used as important outcome measures in HTAs of health technologies. PROMs may allow a further understanding of treatment impact beyond clinical endpoints, but they must be developed with the involvement of beneficiaries of PROMs, such as healthcare providers and patients (7-9). Healthcare providers and patients could benefit from their contribution to PROMs, as they could ultimately prescribe or receive cost-effective treatment, according to evidence collected with PROMs, especially in scenarios where randomized controlled trials (RCTs) are scarce (7). The beneficiaries' involvement is needed, as interdisciplinary advocacy efforts can contribute to the standardization of PROMs and relevant procedures of implementing PROMs (8). Currently, it remains a challenge to engage patients, due to lack of general guidance and engagement strategies (9,10). The IHTAM framework could promote this engagement, by providing general guidance on what beneficiaries can do, and by functioning as a starting point to develop more detailed guidance and engagement strategies specifically suited to PROMs.

As the IHTAM framework could improve interdisciplinary collaboration, it may facilitate the adoption of novel HTA methods in practice and speed up the process of addressing challenges brought by novel health technologies. One example is innovation of HTA methods to promote the use of health technologies related to the application of artificial intelligence (AI). While AI technologies may revolutionize the healthcare system, by automating routine tasks to reduce health-related costs and enhance accessibility of healthcare delivery, their distinct features pose challenges to HTA methods of all aspects (e.g. methods related to data collection, analysis, and decision-making) (11). The issue is further complicated, as a systematic evaluation of the AI technologies may be conducted in a complex real-world context (12). Consequently, the distinctiveness and complexity of AI could hinder the understanding or adoption of the relevant HTA methods by method developers (e.g. researchers), practitioners (e.g. HTA agencies), or beneficiaries (e.g. regulators). As mentioned in the previous two paragraphs, the IHTAM framework could help improve interdisciplinary collaboration and contribute to the development of method-specific guidance and engagement strategies and subsequently be used for guiding the innovation of AI-related HTA methods.

The research in the third chapter demonstrates that the relatively insufficient available time of stakeholders may hinder the whole process of innovation of HTA methods. Given the limited resources (e.g. time), it is difficult for a single stakeholder to

participate in all innovation activities during an innovation process. The roadmap we developed may help to address this conflict, by providing an actionable checklist with 48 items, which defines what stakeholders (i.e. developers, practitioners, and beneficiaries) should do, and possibly when they may hand over their tasks. Additionally, the roadmap defines a pattern of collaboration that may further reduce the complexity of interdisciplinary collaboration. With the roadmap, stakeholders may know where they should take responsibility and when their roles may be taken over.

Moreover, this research identifies a challenge of innovation, related to the order of innovating different HTA methods. We found that the progress of innovating a specific HTA method may depend on the progress of innovating another method. As shown in Case study 1 of Chapter 3, before development of a health economics model, a number of risk prediction models, which function as a part of a health economics model, may be first developed and implemented. In Case study 2, before developing risk prediction models, inputs from stakeholders who develop a health economics model is needed, which could affect the way a risk prediction model is developed and the model transferability. Given the potentially close relationship among different HTA methods, HTA stakeholders may need to identify the order of innovating the HTA methods, and to assess whether and how such order may affect the innovation. Hence, we recommend stakeholders to develop a tailored IHTAM framework for their method to be innovated. The tailored framework could help stakeholders identify additional methods which act as reasons why the target method should be developed, or as factors that influence the ways of method development.

It is also worth noting that the conceptual research of innovating HTA methods as presented in this thesis could contribute to the expansion of an HTA network, by linking the HTA agencies with similar methodological concerns. For example, in Western (e.g. the Netherlands and the UK), Northern (e.g. Sweden), and Southern Europe (e.g. Spain and Italy), HTA agencies and researchers are developing novel methods in response to emerging novel health technologies, such as devices and digital health (13-16). In less developed countries, HTA stakeholders also need to address methodological concerns related to novel health technologies, but focus more on the implementation of existing HTA methods. For example, after a process to appraise digital interventions was established in Europe, researchers recommended Asian HTA agencies (e.g. China) to establish their own process, by learning from European experience (17). As the IHTAM framework provides general guidance on how to identify the needs for a method and subsequently on how to develop and implement a method, the framework could help HTA stakeholders from different agencies or HTA contexts to better understand their complementary capabilities and recognize some potential for collaboration.

## The practice of HTA method innovation: three cases of RWD-related methods

In the following parts of this thesis, we illustrated the process of developing or implementing RWD-related methods. The methods included: (1) tools to assess quality of observational studies (Chapter 4 and 5); (2) tools to assess quality of risk prediction models (Chapter 6); and (3) statistical approaches that merge RCTs and RWD in a (network) meta-analysis (Chapter 7 and 8).

In Chapter 4 and 5, we studied the process of developing or implementing qualitative RWD-related methods. In the fourth chapter, a systematic review was conducted to assess the methodological quality of retrospective observational studies investigating efficacy of diabetes monitoring systems, using the Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) tool. A total of 72 studies were identified, and overall quality was poor with the quality only slightly improving over time.

Our review showed the barriers of applying retrospective observational studies to clinical, regulatory, or HTA decision-making. We found that, given the only slightly increasing trends in quality, the existing efforts to improving quality of retrospective observational studies, such as dissemination of robust quality assessment tools (e.g. ROBINS-I) to researchers, could not fully address relevant concerns about risk of bias (RoB). Though relevant tools were published (18,19), channels were lacking for providing detailed illustration of the tools. Hence, we recommend further research to explore how existing tools were disseminated and the association between dissemination strategies and user adoption.

In this case of applying the ROBINS-I tool to assess the quality of studies using RWD, we experienced that ROBINS-I is generally applicable, but it still has some limitations. For example, it is time-consuming for a pair of reviewers to apply ROBINS-I and to resolve discrepancies on quality judgement (20). Such limitations could consequently reduce the applicability of this tool. Also, we found that limitations of ROBINS-I might not be identified until being applied in practice, so close contact between tool practitioners and developers is necessary. This finding is consistent with a general problem of implementing methods. As mentioned for the IHTAM framework, in the process of implementing a method, developers may explore the implementation practice by the practitioners, otherwise the limitations of a method may not be identified (21). The same applies to the use of the ROBINS-I tool. To further improve this tool, several workshops have been organized and validation studies have been conducted (20,22), but the impact of these efforts is still unknown. Hence, we recommend future research

to investigate the impact of these dissemination and validation efforts, and to improve the ROBINS-I tool continuously.

As the ROBINS-I tool is a qualitative tool related to the use of RWD, it includes technical details that are hard to be understood by practitioners. For example, the tool includes signaling questions with many concepts, such as inception bias, that may not be known by practitioners (23). The difficulty of understanding these technical details then becomes a knowledge gap that would take much time to fill. The IHTAM framework (Part 1), which was informed by design thinking theory and implementation science, has suggested filling this knowledge gap by informing developers to provide technical assistance to practitioners. In the case of the ROBINS-I tool, the existing technical assistance includes a scientific article for elaboration (18), but some other types of assistance may also help, such as interactive question-and-answer programs (24).

To address the complexity of tools identified in the fourth chapter and to facilitate implementation of the tools for assessing quality of studies using RWD, we aimed to summarize and compare signaling questions or criteria of such tools in the fifth chapter. To achieve this, we conducted a targeted search of appraisal tools published from 2002 onwards and a content analysis summarizing quality items from identified tools, using both deductive and inductive coding techniques. Of the 49 identified tools, the RTI Item Bank and the STrengthening the Reporting of OBservational studies in Epidemiology tool (STROBE) were the most comprehensive tools for the assessment of methodological quality and reporting, respectively. However, none of the tools covered all items that are relevant for either methodological quality or reporting. Our research implied that even the most comprehensive tools could be complemented with items from other tools. Therefore, we suggest decision-makers, researchers, and tool developers consider the quality-item level heterogeneity when selecting a tool or identifying a research gap.



From this research, we identified another challenge of innovating HTA methods, i.e., selection of a method that is suitable to a context where needs are identified. Since HTA methods are developed repeatedly, in response to emerging health technologies and the changing needs of various HTA contexts, practitioners from a given context might experience difficulties in the selection of tools that satisfy their own needs. For developers, the lack of information on potentially available methods might cause a waste of resources, if methods that satisfy their similar needs are developed multiple times. This challenge has also been identified in innovating other types of RWD-related methods, such as selection of risk prediction models and tools for measuring patient-reported outcomes (25,26).

To address this challenge, HTA stakeholders could conduct up-to-date reviews and content analyses, just as we did in this study. However, this approach is time-consuming and limited by the difficulty of defining the criteria used for evaluating and selecting the methods. For example, to select tools to judge the quality of studies using RWD, there is an ongoing debate on what criteria may be used (18). Our research tried to collect all criteria used from previous reviews, but still considered the lack of standard criteria a limitation. A potentially necessary step of establishing a set of criteria is to summarize all potential ways of categorizing a type of methods, e.g., based on their aims, functions, and contexts where they can be applied. For example, some tools were developed for assessing methodological quality of studies using RWD, while some other tools were developed for assessing reporting. In addition, some tools, such as the Good ReseArch for Comparative Effectiveness (GRACE) checklist, function as a gateway, to help users have a quick overview of study quality, while some tools, such as ROBINS-I, function as a tool that could provide detailed information on study quality (23,27,28). If all such ways of categorizing a type of HTA method are identified and known by the practitioners, practitioners could select a method that could better satisfy their needs.

In the sixth chapter, we conducted a systematic review to identify models for predicting the risk of coronary heart disease (CHD) in diabetic patients and to assess model quality in terms of RoB and applicability for HTA. Only one of the 23 model studies showed a low RoB in all domains, and no model was fully applicable for health economics modelling. We discovered that most of the major contributors to a high RoB were located in the analysis domain, which was consistent with previous findings (29,30). Also, we learned that model developers mostly did not understand the needs of HTA stakeholders (practitioners and beneficiaries), and we recommend further research to explore the reasons for not understanding these needs. In addition, we emphasized the needs for developing tailored tools for assessing quality of risk prediction models. As the existing tools only address general model applicability, they could not satisfy the needs of HTA stakeholders.

We also noticed that none of the risk prediction models we identified paid attention to the implementation of these models in HTA practice. While the use of risk prediction models in clinical settings is frequently mentioned, developers (e.g. researchers) seemed to be unaware of the heterogeneous needs from different disciplines (e.g. HTA and clinical practice) and contexts, e.g., in terms of therapeutic and geographic areas. This finding was consistent with a previous review which highlighted the necessity to provide guidance to standardize the model implementation, so risk prediction models could be adequately used (31). To provide guidance, constructing a tailored

IHTAM framework, which takes into account issues related to risk prediction models, may be a good way forward. In this tailored framework, the case-specific definitions of developers, practitioners, and beneficiaries can be provided. For example, model developers should have knowledge of which variables are the most likely to be risk factors (32). Hence, not only academic researchers, but also clinicians may be methods developers.

In the third part of this thesis (chapters 7 and 8), we studied the process of developing and implementing another quantitative HTA method that promoted the use of RWD in HTA settings, i.e., statistical approaches to merge data from RCTs and RWD. In the seventh chapter, we conducted three network meta-analyses to investigate diabetes monitoring systems with insulin delivery in type 1 diabetes mellitus (T1DM) patients, with studies using RWD only, RCTs only or both as evidence, and investigated whether the estimated efficacy differed.

Our study showed that the efficacy estimated from pooling data from NRSs generally differed from the estimates obtained based on pooling of RCTs data, but the results were not statistically significant for NRSs. In contrast, results estimated from RCT data only and from combined evidence were mostly similar, as the NMA was dominated by RCTs. Also, changing the NRSs' weight relative to RCTs, especially for those NRSs with serious risk of bias as defined by ROBINS-I, could have a large impact on the estimated efficacy.

The findings on whether the RCTs and NRSs provided consistent pooled estimates differed among the previous studies across disease fields. For example, Hong et at. (2021) searched PubMed and Embase for systematic literature reviews which reported relative treatment effects of pharmaceuticals from both observational studies and RCTs. They also analyzed pairs of pooled effect estimates from RCTs and observational studies, and detected significant differences in only 20% of the 74 pairs (33). In our study, the relatively consistent results from RCTs and combined evidence may be explained by the small number of NRSs and the power prior approach which assigned NRSs a low weight. We also found that downweighing NRSs based on RoB, complemented by scenario analyses to investigate relevant impacts, might be a good strategy when incorporating NRSs into an NMA.



In this study, we confirmed the finding of the fifth chapter that selection of a method that is suitable to a context where needs are identified was a challenge of innovating HTA methods. Currently, there are still debates on whether the power prior approach is the most appropriate method for merging data from RCTs and RWD in an NMA.

According to an overview published in 2020, the power prior approach was suitable to scenarios where RCTs were the gold-standard evidence but might lead to biased estimates on efficacy where RCTs were rare (34). Some other available approaches, such as hierarchical modelling, might complement the power prior approach, but the most appropriate scenarios where it could be used remains uncertain (34). The difficulty of the selection of the best approach was further complicated by technical issues, as the same approach (e.g. hierarchical modelling) written in R or the BUGs language by different authors might have different assumptions, codes, or even mathematical algorithms listed behind (35,36).

In addition, while these statistical approaches that merge data from RCTs and RWD were promising, their transferability to the HTA context was limited. From one aspect, we noticed a lack of consistency in concepts used by the approach developers in their publications or manuals to describe how a statistical approach could be used. The recently published original studies for developing the power prior or the hierarchical modelling approach and relevant reviews often used different wording to describe the same approach, the approach features, or the process of developing an approach (37,38). For example, the power prior approach was sometimes referred to as "as prior information" or "using informative priors", and the same type of assumptions were often presented with different mathematical expressions (37,38). Consequently, HTA stakeholders might feel these methods are hard to understand, let alone adopt these approaches. One solution may be making a tutorial on how these approaches could be implemented by HTA stakeholders, in which relevant concepts are clearly defined and technical details are explained with examples. On the other hand, in the HTA context, (network) meta-analyses are normally used as evidence for assessing the effectiveness and cost-effectiveness (39). The existing validation studies of these statistical approaches, either qualitative or quantitative, focused on the approach impact on the clinical effectiveness (40,41). While these validation studies are important, further impact analysis of these statistical approaches on the cost-effectiveness may help encourage the engagement of practitioners from the HTA context (e.g. HTA agencies or researchers who conduct NMAs). This increased engagement may help approach developers understand the needs and facilitate adoption of the approaches.

In the eighth chapter, we applied the three approaches that could merge data from RCTs and RWD in a meta-analysis (i.e. naïve pooling, power prior, and hierarchical modelling), using the Crossnma R package that was developed as part of our HTx project. We aimed to compare whether the pooled estimates were consistent. The three approaches were applied with Bayesian random-effects meta-regression models in four cases: two case studies on myelodysplastic syndromes and two on diabetes. The study

showed that the level of consistency of results obtained using the three approaches varied in the four cases. Also, the hierarchical modelling approach resulted in much larger uncertainty than the power prior approach. Our study implied that the Crossnma package provided a user-friendly way of synthesizing evidence from RCTs and studies using RWD, and it had potential to be applied by practitioners without a statistical background. Also, the implementation of statistical approaches that synthesize data from RCTs and studies using RWD could enhance the complementation of the two data sources, thereby overcoming some of their limitations.

From this research we noticed that the selection of a statistical approach that merges data from RCTs and RWD in a meta-analysis could be a challenge. In Chapter 5, we highlighted the importance of considering both risk of bias and reporting when selecting an appropriate tool to assess quality of studies using RWD. In the case of selecting a statistical approach that merges RWD and data from RCTs, the relevant criteria should not only include risk of bias or reporting, but also statistical performance of the approach (e.g. accuracy of obtained pooled estimates). This also applies to other quantitative methods. For example, to select a risk prediction model, practitioners should not only assess model quality, in terms of risk of bias, but also evaluate model discrimination and calibration with relevant statistical tests, such as the concordance statistic (42,43). Given the necessity of assessing multiple aspects when selecting a quantitative method comprehensive guidelines should be developed with instructions on how to validate a method. Moreover, since existing guidelines may focus only on some quality concerns, such as risk of bias and statistical performance, and may not consider the needs of HTA stakeholders (e.g. user-friendliness of applying a statistical approach) (44,45), we suggest involving HTA stakeholders (e.g. researchers who gather HTA evidence with meta-analyses) in evaluating and improving these guidelines. Also, in this case, a tailored IHTAM framework for innovating a certain HTA method might help guideline developers identify stakeholders who can provide insights.

The Crossnma package that we used for applying the statistical approaches provides a simple way of comparing several statistical approaches for the purpose of validation, an important subphase of HTA method innovation according to our IHTAM framework. In contrast, before the Crossnma package was available, the statistical approaches might only be applied or compared if practitioners were experienced in coding, e.g., with the BUGs language (36,37). The idea of improving the method's user-friendliness by standardizing the way of implementing multiple methods can be conveyed to other types of methods. For example, in Chapter 5, the existence of more than 40 tools to assess the quality of studies using RWD poses a challenge for tool selection and implementation. As prompted by the research in Chapter 8, an interactive tool that

lists the existing quality assessment tools, with functions to filter the tools according to the tool features, could simplify the ways of comparing and implementing these tools.

### Implications related to the use of RWD in the HTA context

In summary, our findings from Chapter 4-8 can help improve the use of RWD in the HTA context, mainly in two aspects. First, the results of this thesis could contribute to the development of policies for use of RWD in HTA. One aim of policies made by the HTA and regulatory agencies is to help HTA stakeholders address methodological challenges and select an appropriate method (45). For example, NICE and ZIN have developed policies to clarify evidence requirements for reimbursements of pharmaceuticals across Europe, including methods needed for evidence collection and analysis (46,47). Our research on innovation of tools to assess quality of studies using RWD and approaches to merge data from RCTs and RWD could serve as a starting point for updating the current HTA policies on evidence collection, as we provided an overview of the existing methods (Chapter 5 and 8) and illustrated how they could be used (Chapter 4, 5 and 7). Similarly, the research presented in this thesis could provide input to the policies on health economics analyses, as our research on risk prediction models (Chapter 6) helps address methodological concerns on how a risk prediction model could be incorporated into health economics modelling.

Second, the HTA methods described in our research may help address some challenges of using RWD for the purpose of HTA. For example, according to a review published in 2021, research on the effectiveness of treatments using electronic patient records was often hindered by the lack of information on comparative treatments (48). This limitation was also observed in Chapter 3, in which we found that approximately one-third of retrospective observational studies investigating efficacy of diabetes monitoring systems lacked a comparator. This challenge could be partly addressed by the approaches to merge data from RCTs and RWD. With the Crossnma package, the single-treatment groups from RWD could be incorporated in a network meta-analysis, which could be ultimately used as evidence for the HTA decision-making (38).

Another challenge of using RWD in the HTA context is the difficulty of evaluating quality concerns related to RWD. As RWD have higher risk of bias than RCTs, evidence obtained from RWD should be first downweighed (e.g. based on overall RoB) before they are used for HTA decision-making, otherwise they may not be trusted by HTA stakeholders (2,49). However, HTA stakeholders may face difficulty in determining the overall RoB, if they lack an overview of all relevant quality concerns or expertise in how these concerns should be evaluated (18,19). This challenge can be gradually overcome, if continuous efforts for developing, selecting, and implementing robust quality

assessment tools are made by stakeholders both within and beyond the HTA context (e.g. using RWD for the regulatory purpose), and HTA stakeholders are informed about any progress promptly. This thesis could help HTA stakeholders track the progress of addressing RWD-related quality concerns, as we identified all recently developed tools, summarized and compared their signalling questions, and provided suggestions about how these tools could be selected.

## Potential solutions to the challenges of HTA method innovation

In the introduction part of this thesis (Chapter 1), we mentioned that the main problems related to HTA method development included the lack of a clear overview of the needs from HTA stakeholders and the limitation of resources, such as available time and high-quality data. An essential problem related to HTA method implementation was the lack of expertise of skills to implement HTA methods. In Chapter 2-8, we further conducted conceptual research to address the challenges related to the development and implementation of HTA methods and illustrated how this conceptual research could help address these challenges, using RWD-related HTA methods as the cases. In this section, we will discuss potential solutions to the above-mentioned challenges, by summarizing findings of the previous chapters.

Before discussing the solutions, we have to mention that, in the following paragraphs, we refer to HTA stakeholders as developers, practitioners, or beneficiaries, rather than as the classical stakeholder groups, such as researchers, HTA agencies, and healthcare professionals. The reason is that developers, practitioners, and beneficiaries are more suited to the description of general issues related to method innovation, and their implications depend on the method being innovated. For example, practitioners of a risk prediction model can be healthcare professionals and researchers, while practitioners of a tool for assessing quality of studies using RWD can be HTA agencies.



### *The lack of a clear overview of the needs from HTA stakeholders*

As mentioned in Chapter 1, method developers need approaches that facilitate their understanding on the needs from other HTA stakeholders, i.e., practitioners and beneficiaries. To start with addressing this problem, method developers should acquire insight in limitations of current HTA processes, e.g., by gaining feedback from practitioners who used traditional methods and beneficiaries who are affected by them. Some of the other suggestions, as mentioned in the IHTAM framework, include picturing what future HTA processes looks like and evaluating heterogeneity of contexts where research gaps are identified. In addition, we recommend developers to follow the advice given in the IHTAM framework about how to identify the needs, such as to conduct update-to-date reviews, surveys, interviews, or brainstorm sessions. The

reason is that the needs for an HTA method may vary across contexts, and needs may change due to emerging novel health technologies.

In order to provide a clear overview of the needs from HTA stakeholders, we also recommend method developers to invite potential method practitioners and future beneficiaries to engage in the process of method development. To encourage practitioners and beneficiaries to engage, we recommend method developers to clearly report how technical details of a method could help satisfy the needs. For example, in the process of developing a tool for assessing quality of cohort studies (i.e. RWD studies in which a group of people with a common characteristic is followed over time), developers should inform practitioners and beneficiaries of the reasons why criteria of this tool are relevant to quality of cohort studies, and the ways of judging the overall study quality using these criteria. With information about all technical details, practitioners and beneficiaries could provide timely feedback and point out potential concerns that limit the method's applicability or transferability.

***The limitation of resources, such as available time and high-quality data***

To resolve the conflicts between a long project duration and the relatively insufficient available time of stakeholders, we recommend HTA stakeholders, i.e. developers, practitioners, and beneficiaries, to first list all potential actions needed for developing and implementing a method. Some example actions include evaluating the necessity of developing a novel method; setting priorities for needs (e.g. those identified from various contexts) that a method addresses, given limited resources; and developing a implementation strategy for guiding the resources needed for conducting and monitoring the implementation and for motivating potential practitioners to adopt the novel HTA method.

To address the limitations due to the lack of some other types of resources (e.g. high-quality data and knowledge), we recommend HTA stakeholders from different HTA contexts to better understand their complementary capabilities of developing or implementing an HTA method, and to collaborate in a larger HTA network. To recognize potential for collaboration, we, again, refer HTA stakeholders to the general guidance provided by the IHTAM framework roadmap in Chapter 3. As the roadmap provides an overview of all phases of innovation, a stakeholder may better understand where they could make contributions and where they need a collaborator.

***The lack of expertise or skills to implement a method in the HTA context***

The lack of expertise or skills to implement a method in the HTA context is a main challenge of HTA method implementation. To address this challenge, we recommend

developers to inform practitioners and beneficiaries of the contexts (e.g. in terms of geography or therapeutic areas) where using a method is appropriate. The reason is that, as discussed in Chapter 5 and 8, practitioners from a given context might experience difficulties in the selection of methods that satisfy their own needs. To judge in which context implementing a method is appropriate, method developers may first refer to aims, functions, and internal and external validity of a method. Method developers may also share their initial judgement with practitioners and beneficiaries, and check whether their suggestions for the method selection can be adopted or need to be further improved. Finally, a guideline, manual, or interactive tool that guides the method selection can be created, then disseminated to all HTA stakeholders.

## Study limitations

The research described in the thesis mostly focuses on the application of innovative HTA methods but less on developing new HTA methods. For instance, in Chapter 4 and 5, we illustrated how the tools to assess the quality of studies using RWD were used and selected and in Chapter 6, we mainly discussed the challenges of applying risk prediction models for HTA purposes. Additionally, in Chapter 7 and 8, we applied the statistical approaches to merge data from RCTs and RWD in (network) meta-analyses. One reason for the absence of research on HTA method development in this thesis is the long duration of a method innovation process, which often involves multiple stakeholders from various institutes. In the HTx project, the tasks on (HTA) method development were mainly assigned to researchers of several institutes (e.g. Technical University of Madrid), while our team (Utrecht University) mostly focused on the method application. Given the long innovation process and the complexity of project management, the conceptual research on HTA method innovation and the improvement of the IHTAM framework need joint and continuous effort. Another limitation of our mostly conceptual research was that we were only able to apply three methods (e.g. risk prediction models) in this thesis for the illustration purpose and were not able to involve all classical HTA stakeholder groups (e.g. healthcare providers, payers, and industry) in the method innovation. Consequently, the applicability of our findings and recommendations in response to the general problems of HTA methods innovation (e.g. using the IHTAM framework to address the complexity of resource management) will need more research using case studies in diverse settings. A final limitation was that we did not cover the full spectrum of relevant issues concerning the use of RWD for the purpose of HTA. For example, if local RWD is lacking, transferability of RWD becomes a topic with technical challenges, and this requires innovation of novel methods (50). In this thesis, we mainly focused on the methods for assessing

the quality of studies using RWD and methods for synthesizing RWD (Chapter 4-5 and Chapter 7-8). We believe that the guidance provided in this thesis regarding HTA methods innovation could be applied to all types of methods development. Therefore, we recommend that the IHTAM framework is applied to future research that addresses additional RWD-related challenges for HTA.

## Future opportunities

In order to implement and test the solutions proposed in this thesis to the general problems related to developing and implementing HTA methods, all groups of classical HTA stakeholders (e.g. HTA agencies, researchers, and industry) from different contexts (e.g. countries) need to make joint and continuous effort. Some future opportunities to address the problems related to method development and implementation are listed as follows.

First, we advise HTA stakeholders participating in a future project addressing HTA methodological concerns to not only develop a robust method or to implement a method successfully, but also to identify and summarize additional challenges of method innovation that occur within their project. Although this thesis has listed and discussed some challenges of method innovation, we may not have been able to identify all potential challenges in a thesis investigating only three types of methods (e.g. statistical approaches to merge data from RCTs and RWD). Also, we advise HTA stakeholders to propose potential solutions to challenges of innovation and to consider these solutions as a part of their project outcomes. The reason is that the challenges of method innovation which they identify can occur in developing or implementing various types of methods, and solutions they propose may contribute to a successful method innovation and improve method quality across HTA contexts.

Second, we advise future projects addressing HTA methodological concerns to focus more on challenges and solutions related to the successful implementation of a method. For example, a future project may investigate how a type of method can be selected, based on the stakeholders' needs. As shown in Chapter 4-8, some projects addressing HTA methodological concerns focus relatively more on developing a method. In these projects, the method applicability is only tested in several cases studies, evaluated by a small group of stakeholders, and the method performance is monitored in a short period of time. In order to thoroughly investigate the implementation issues of a method, different projects investigating HTA methodological concerns should maintain continuity in innovating a type of method. Also, we advise that practitioners

and beneficiaries, e.g., those with non-academic backgrounds, should play a major role in projects investigating issues related to method implementation, while developers can provide technical assistance.

Third, future research may offer a comprehensive overview of linkage of issues related to innovating different types of HTA methods and issue an order of innovating the methods. The reason is that, as mentioned in Chapter 3 of this thesis, the progress of innovating a specific HTA method may depend on the progress of innovating another, and following a specific order might lead to a more efficient innovation process. For example, a risk prediction model should be first developed and implemented, before developing a health economics model, and the validity of a health economics model partly depends on the validity of a risk prediction model. The existing reviews that summarize methodological concerns related to a certain type of method, such as those related to RWD methods, can already be quite comprehensive (51). However, future research is needed to investigate whether these concerns should be prioritized or temporarily put on hold, depending on the progress made for addressing methodological concerns of another type of methods. One suggestion for offering such a comprehensive overview is to conduct conceptual research to identify the relationships between innovation progresses and methodological concerns, through synthesizing multiple reviews, each of which summarizes methodological issues related to a certain type of method.

## Conclusion



The innovation, i.e. development and implementation, of HTA methods is a natural process needed for improving HTA quality and relevance, in response to emerging novel health technologies and the variety in HTA practices. However, the innovation process involves several challenges, which may be resolved by the conceptual framework and associated tools to guide the process of HTA method innovation this thesis provides. The use of the framework may lead to increased awareness that not only researchers, but also all HTA stakeholders (e.g. patients, clinician and health technology developers) have the potential to contribute to the development or implementation of HTA methods. It may also support a more standardized and structured process of innovating HTA methods. Finally, HTA stakeholders from multiple disciplines may find better ways of collaboration to innovate HTA methods, by following and adjusting the framework and tools provided in this thesis. In doing so, new challenges in the HTA context may be addressed more efficiently and consistently.

## Author contribution

LJ wrote and edited the introduction. The supervisory team provided feedback throughout the process and approved the final version.

# References


1. Hogervorst MA, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. Reported challenges in health technology assessment of complex health technologies. Value Health. 2022 Jun 1;25(6):992-1001.
2. Oortwijn W, Sampietro-Colom L, Trowman R. How to deal with the inevitable: generating real-world data and using real-world evidence for HTA purposes–from theory to action. Int J Technol Assess Health Care. 2019;35(4):346-50.
3. Lou J, Sarin KC, Toh KY, et al. Real-world data for health technology assessment for reimbursement decisions in Asia: current landscape and a way forward. Int J Technol Assess Health Care. 2020 Oct;36(5):474-80.
4. Garrido MV, Gerhardus A, Røttingen JA, Busse R. Developing health technology assessment to address health care system needs. Health Policy (New York). 2010 Mar 1;94(3):196-202.
5. Stevens A, Milne R, Burls A. Health technology assessment: history and demand. J Public Health. 2003 Jun 1;25(2):98-101.
6. Banta D. The development of health technology assessment. Health policy. 2003 Feb 1;63(2):121-32.
7. Whittal A, Meregaglia M, Nicod E. The use of patient-reported outcome measures in rare diseases and implications for health technology assessment. Patient : patient-centered outcomes research. 2021 Sep;14:485-503.
8. Zannad F, Alikhaani J, Alikhaani S, et al. Patient-reported outcome measures and patient engagement in heart failure clinical trials: multi-stakeholder perspectives. Eur J Heart Fail. 2023 Apr;25(4):478-87.
9. Haywood K, Brett J, Salek S, et al. Patient and public engagement in health-related quality of life and patient-reported outcomes research: what is important and why should we care? Findings from the first ISOQOL patient engagement symposium. Qual Life Res. 2015 May;24:1069-76.
10. McNeill M, Noyek S, Engeda E, Fayed N. Assessing the engagement of children and families in selecting patient-reported outcomes (PROs) and developing their measures: a systematic review. Qual Life Res. 2021 Apr;30:983-95.
11. Bélisle-Pipon JC, Couture V, Roy MC, Ganache I, Goetghebeur M, Cohen IG. What makes artificial intelligence exceptional in health technology assessment?. Front Artif Intell. 2021 Nov 2;4:153.
12. Alami H, Lehoux P, Auclair Y, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res. 2020 Jul 7;22(7):e17707.
13. Akehurst RL, Abadie E, Renaudin N, Sarkozy F. Variation in health technology assessment and reimbursement processes in Europe. Value Health. 2017 Jan 1;20(1):67-76.
14. Fuchs S, Olberg B, Panteli D, Perleth M, Busse R. HTA of medical devices: Challenges and ideas for the future from a European perspective. Health Policy. 2017 Mar 1;121(3):215-29.
15. Fuchs S, Olberg B, Panteli D, Busse R. Health technology assessment of medical devices in Europe: processes, practices, and methods. Int J Technol Assess Health Care. 2016;32(4):246-55.
16. Srivastava D, Henschke C, Virtanen L, et al. Promoting the systematic use of real-world data and real-world evidence for digital health technologies across Europe: a consensus framework. Health Econ Policy Law. 2023 Oct;18(4):395-410.
17. Barzey V, Brennan J, Tutt S, Cope N. Innovative Access Pathways for Digital Healthcare Solutions: Learnings From a European Analogue Analysis for Asian HTA Bodies. Value Health. 2018 Sep 1;21:S73.
18. Quigley JM, Thompson JC, Halfpenny NJ, Scott DA. Critical appraisal of nonrandomized studies—a review of recommended and commonly used tools. J Eval Clin Pract. 2019 Feb;25(1):44-52.
19. D'Andrea E, Vinals L, Patorno E, et al. How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools. BMJ open. 2021 Mar 1;11(3):e043961.

20. Jeyaraman MM, Rabbani R, Al-Yousif N, et al. Inter-rater reliability and concurrent validity of ROBINS-I: protocol for a cross-sectional study. Systematic reviews. 2020 Dec;9:1-2.
21. Jiu L, Hogervorst MA, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. Understanding innovation of health technology assessment methods: the IHTAM framework. Int J Technol Assess Health Care. 2022;38(1):e16.
22. Jeyaraman MM, Rabbani R, Copstein L, et al. Methodologically rigorous risk of bias tools for nonrandomized studies had low reliability and high evaluator burden. J Clin Epidemiol. 2020 Dec 1;128:140-7.
23. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016 Oct 12;355.
24. Wang Z, Taylor K, Allman-Farinelli M, et al. A systematic review: Tools for assessing methodological quality of human observational studies. MedRxiv. (Preprint). May 21, 2019. https://doi.org/10.31222/osf.io/pnqmy
25. Li X, Li F, Wang J, van Giessen A, Feenstra TL. Prediction of complications in health economic models of type 2 diabetes: a review of methods used. Acta Diabetol. 2023 Jul;60(7):861-79.
26. Valderas JM, Ferrer M, Mendívil J, et al. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. Value Health. 2008 Jul;11(4):700-8.
27. Dreyer NA, Bryant A, Velentgas P. The GRACE checklist: a validated assessment tool for high quality observational studies of comparative effectiveness. J Manag Care Spec Pharm. 2016 Oct;22(10):1107-13.
28. Dreyer NA, Velentgas P, Westrich K, Dubois R. The GRACE checklist for rating the quality of observational studies of comparative effectiveness: a tale of hope and caution. J Manag Care Spec Pharm. 2014 Mar;20(3):301-8.
29. Haider S, Sadiq SN, Moore D, Price MJ, Nirantharakumar K. Prognostic prediction models for diabetic retinopathy progression: a systematic review. Eye. 2019 May;33(5):702-13.
30. Van der Heijden AA, Nijpels G, Badloe F, et al. Prediction models for development of retinopathy in people with type 2 diabetes: systematic review and external validation in a Dutch primary care setting. Diabetologia. 2020; 63(6):1110-1119.
31. van Giessen A, Peters J, Wilcher B, et al. Systematic review of health economic impact evaluations of risk prediction models: stop developing, start evaluating. Value Health. 2017 Apr 1;20(4):718-26.
32. Rosenberg AL. Recent innovations in intensive care unit risk-prediction models. Curr Opin Crit Care. 2002 Aug 1;8(4):321-30.
33. Hong YD, Jansen JP, Guerino J, et al. Comparative effectiveness and safety of pharmaceuticals assessed in observational studies compared with randomized controlled trials. BMC Med 2021 Dec;19(1):1-5.
34. Sarri G, Patorno E, Yuan H, et al. Framework for the synthesis of non-randomised studies and randomised controlled trials: a guidance on conducting a systematic review and meta-analysis for healthcare decision making. BMJ Evid Based Med. 2022 Apr 1;27(2):109-19.
35. Verde PE. A bias-corrected meta-analysis model for combining, studies of different types and quality. Biom J. 2021 Feb;63(2):406-22.
36. McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride JE. The importance of adjusting for potential confounders in Bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominal aortic aneurysms. BMC Med Res Methodol. 2010 Dec;10:1-9.
37. Schmitz S, Adams R, Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. Stat Med. 2013 Jul 30;32(17):2935-49.

38. Efthimiou O, Mavridis D, Debray TP, et al. Combining randomized and non-randomized evidence in network meta-analysis. Stat Med. 2017 Apr 15;36(8):1210-26.

39. Chaiyakunapruk N, Saokaew S, Sruamsiri R, Dilokthornsakul P. Systematic review and network meta-analysis in health technology assessment. J Med Assoc Thai. 2014 May 1;97:S33-42.

40. Yao M, Wang Y, Mei F, Zou K, Li L, Sun X. Methods for the Inclusion of Real-World Evidence in a Rare Events Meta-Analysis of Randomized Controlled Trials. J Clin Med. 2023 Feb 20;12(4):1690.

41. Jenkins DA, Hussein H, Martina R, Dequen-O'Byrne P, Abrams KR, Bujkiewicz S. Methods for the inclusion of real-world evidence in network meta-analysis. BMC Med Res Methodol. 2021 Dec;21(1):1-9.

42. Pencina MJ, D'Agostino RB. Evaluating discrimination of risk prediction models: the C statistic. Jama. 2015 Sep 8;314(10):1063-4.

43. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making. 2015 Feb;35(2):162-9.

44. Moons KG, Altman DG, Reitsma JB, Collins GS. New guideline for the reporting of studies developing, validating, or updating a multivariable clinical prediction model: the TRIPOD statement. Adv Anat Pathol. 2015 Sep 1;22(5):303-5.

45. Moons KG, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. 2019 Jan 1;170(1):W1-33.

46. Makady A, Ten Ham R, de Boer A, Hillege H, Klungel O, Goettsch W. Policies for use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies. Value Health. 2017 Apr 1;20(4):520-32.

47. Taylor J, Patrick H, Lyratzopoulos G, Campbell B. Methodological challenges in evaluating the value of registers. Int J Technol Assess Health Care. 2014 Jan;30(1):28-33.

48. Nederland, Z. Beoordeling stand van de wetenschap en praktijk. Diemen: Zorginstituut Nederland. 2015.

49. Griffiths EA, Vadlamudi NK. Not ready for the real world? The role of non-RCT evidence in health technology assessment. Value Health. 2016 May 1;19(3):A286.

50. Jaksa A, Arena PJ, Chan KK, Ben-Joseph RH, Jónsson P, Campbell UB. Transferability of real-world data across borders for regulatory and health technology assessment decision-making. Front Med (Lausanne). 2022 Nov 16;9:1073678.

51. Jaksa A, Wu J, Jónsson P, Eichler HG, Vititoe S, Gatto NM. Organized structure of real-world evidence best practices: moving from fragmented recommendations to comprehensive guidance. J Comp Eff Res. 2021 Jun;10(9):711-31.

## Summary

Health technology assessment (HTA) is a process of using explicit methods to determine the value of a health technology at different points in its lifecycle. HTA aims to contribute to an efficient and equitable health system, but the quality and relevance of HTA methods is often discussed, for instance on the question of whether HTA methods are appropriate. Based on evidence obtained and synthesized with inappropriate HTA methods, stakeholders could make sub-optimal decisions, e.g., on reimbursement and pricing of new health technologies. To increase the availability of appropriate HTA methods, HTA methods have been repeatedly developed and implemented, since HTA became an important element of healthcare systems in the 1980s. To further clarify the concept "HTA methods", in Chapter 1 we summarized the ways "HTA methods" were mentioned by HTA agencies and in research projects and scientific articles. In summary, the concept embraces the full scope of an HTA process, including evidence collection, evidence appraisal, decision-making, and monitoring. Also, HTA methods can be qualitative in nature and may include frameworks, guidelines, and checklists, or quantitative models and statistical approaches.

Although many new HTA methods have appeared over time, the development and implementation of those methods have been perceived as suboptimal. One type of problem related to developing HTA methods is the lack of a clear overview of the needs of HTA stakeholders. Although the importance of understanding the needs from HTA stakeholders has been increasingly recognized in the HTA field, the approaches that facilitate the understanding by method developers are still lacking. Another type of problem related to HTA method development is the limitation of time, high-quality data, and knowledge across research disciplines. Additionally, HTA method development is limited by the lack of theory for understanding how to develop HTA methods. As no single comprehensive framework for understanding innovation challenges in highly complex research exists, the similarities and differences of research that involve complex innovation activities should be further explored by HTA stakeholders. Regarding HTA method implementation, one problem is that method practitioners may lack expertise on the HTA methods. Moreover, the diversity of methods, each of which involve specific knowledge on how to use the method, could even further complicate the problem of method implementation. Another problem is that HTA stakeholders may lack collaboration skills that enable the use of different research methods. In the case of implementing a method which engages a large and diverse stakeholder group, it is necessary to refine methodology to synthesize conflicting viewpoints or potentially missing stakeholder perspective.

One potential solution to the above-mentioned general problems is to establish and illustrate a pattern of identifying the stakeholders' needs and facilitating stakeholder collaboration, throughout the innovation process. Previous research has built some foundations, by developing guidance for developing some types of HTA methods. However, these studies have several limitations, such as the lack of general guidance for understanding why a method should be developed. Therefore, further research is needed to provide a general guidance on how HTA methods should be innovated and to compare the patterns of innovation among the different types of methods. In addition to illustrating the general problems, Chapter 1 presents the needs for methods to promote the use of real-world data (RWD) in HTA settings. Since RWD can complement randomized-controlled trials (RCTs), especially when RCTs are scarce or infeasible to conduct, there is a growing appetite to use RWD in HTA. However, the usefulness of RWD is often questioned due to quality concerns, such as risk of bias. To facilitate the use of RWD, these quality concerns need to be evaluated and then addressed with appropriate methods. However, considering the emergence of novel health technologies and the variety of settings where HTA is conducted, as mentioned above, the existing methods for evaluating or addressing RWD quality concerns may not satisfy all needs of HTA stakeholders, so novel methods need to be repeatedly developed and implemented. Therefore, this thesis aimed to address these general problems, by conducting relevant conceptual research, and by illustrating how the conceptual research could help address the problems, using RWD-related HTA methods as the cases.

To begin with, Chapter 2 presents a framework with two functions: to illustrate a generic innovation process that is applicable to all types of HTA methods and to illustrate how different HTA stakeholder groups can engage dynamically and collaborate effectively throughout the innovation process. The framework was constructed based on twenty documents on innovating HTA frameworks and fourteen guidelines from three scientific disciplines. It includes a generic innovation process consisting of three phases (“Identification”, “Development”, “Implementation”) and nine subphases. In the framework, three roles that HTA stakeholders can play in innovation (“Developers”, “Practitioners”, “Beneficiaries”) are defined and a process on how the stakeholders innovate HTA methods is included. This framework visualizes systematically which elements and stakeholders are important to the development and implementation of novel HTA methods.

In Chapter 3, we further explored framework validity in three cases of method innovation that are part of the HTx project. The results indicated that the framework provided a structural way of deliberation and helped to improve collaboration

among HTA stakeholders. However, framework applicability could be improved if it is complemented by a roadmap with a loop structure to provide tailored guidance for different cases, and with items to elaborate actions to be taken by stakeholders. Accordingly, a forty-eight-item roadmap was developed with a loop structure and actionable items, which could complement the framework, and might provide HTA stakeholders with tailored guidance on developing new methods.

In the following chapters, we investigated specific research questions related to development or implementation of qualitative and quantitative HTA methods using RWD. More specifically, in Chapter 4,5, and 6, we focused on methods used for assessing quality of studies using RWD; in Chapter 7 and 8, we focused on methods used for merging RCTs and RWD in (network) meta-analyses. In Chapter 4, we conducted a systematic review to assess the methodological quality of retrospective observational studies investigating effects of diabetes monitoring systems, and to explore the trend in quality over time. The results indicated that the overall methodological quality was quite low, as 61 (85%) studies were graded as facing critical or serious risk of bias. Also, the overall methodological quality did not substantially improve over time. The major contributors to low quality included not adequately controlling for confounding, missing data, and selective reporting of the results. Thus, clinical, regulatory, or HTA decision-makers may need strategies to effectively exploit these suboptimal studies. Also, to further improve study quality extra efforts may be needed such as guiding the tool selection regarding quality improvement in the tools.

In Chapter 5, we further identified existing appraisal tools for non-randomized studies of interventions, and compared criteria the tools provided at the quality-item level. Our study identified 49 tools and showed that overall the Research Triangle Institute (RTI) Item Bank and the STrengthening the Reporting of OBservational studies in Epidemiology Checklists (STROBE) were most comprehensive, with the highest number of items addressed and sufficiently described, respectively, on methodological quality and reporting. However, none of the tools addressed concerns in all items, not even briefly. The items least addressed for methodological quality included outcome selection, outcome definition, and ethical approval, and for reporting included intervention selection, intervention measurement, and length of follow-up. The results indicated that most of the appraisal tools had their own strengths, but none of them could address all quality concerns relevant to non-randomized studies of interventions. Even the most comprehensive tools can be complemented by several tools. Therefore, we suggest decision-makers and researchers consider the quality-item level heterogeneity, when selecting a tool.

In Chapter 6, we conducted a literature review to identify models for predicting the risk of coronary heart diseases in patients with diabetes, and to assess model quality in terms of risk of bias and applicability for the purpose of HTA. Of the 26 model studies and 30 models identified, only one model study showed low risk of bias in all domains, and no model was fully applicable for HTA. We advise that, to develop future models, the needs from HTA stakeholders, especially regarding health economics modelling, and the existing quality appraisal tools should be taken into account. Moreover, since general model applicability is not informative for HTA, novel adapted tools may need to be developed.

In Chapter 7, we conducted parallel network meta-analyses investigating diabetes monitoring systems with insulin delivery in patients with type 1 diabetes, using non-randomized studies (NRSs), RCTs, or both as evidence, and investigated whether the estimated efficacy differed. The study showed that RCTs belonged to two separate networks, and they were connected to one network after NRSs were incorporated. NRSs, after being downweighed, could merge and extend the intervention networks of RCTs. In addition, the efficacy and rankings estimated from NRSs differed from those from RCTs, but results were not statistically significant. In contrast, results from RCTs and combined evidence were mostly similar. Additionally, changing the weight of NRSs relative to RCTs, especially for those with serious risk of bias, impacted the estimated efficacy greatly with statistical significance.

In Chapter 8, we applied an R package (i.e. Crossnma) to cases on myelodysplastic syndromes or diabetes, to explore whether the effect estimates obtained from the three statistical approaches supporting synthesis of RCTs and NRSs (i.e. naïve pooling, power prior, and hierarchical modelling) were consistent. The study showed that the power prior approach was more reliable than the naive pooling and hierarchical modelling approach for synthesizing evidence from RCTs and NRSs in a meta-analysis, but none of the approaches could guarantee the accuracy of pooled estimates when the absolute number or proportion of RCTs was small.



Based on the findings from previous chapters, in Chapter 9, we discussed the implications related to the use of RWD in the HTA context. Our findings from Chapter 4-8 can help improve the use of RWD mainly in two aspects. First, the results could contribute to the development of policies for use of RWD in HTA. One aim of policies made by the HTA and regulatory agencies is to help HTA stakeholders address methodological challenges and select an appropriate method. Our research on innovation of tools to assess quality of studies using RWD and approaches to merge data from RCTs and RWD could serve as a starting point for updating the current

HTA policies on evidence collection, as we provided an overview of the existing methods (Chapter 5 and 8) and illustrated how they could be used (Chapter 4, 5 and 7). Similarly, the research presented in this thesis could provide input to the policies on health economics analyses, as our research on risk prediction models (Chapter 6) helps address methodological concerns on how a risk prediction model could be incorporated into health economics modelling. Second, the HTA methods described in our research may help address some challenges of using RWD for the purpose of HTA. For example, one challenge of using RWD in the HTA context is the difficulty of evaluating quality concerns related to RWD. This challenge can be gradually overcome, if continuous efforts for developing, selecting, and implementing robust quality assessment tools are made by stakeholders both within and beyond the HTA context (e.g. using RWD for a regulatory purpose), and HTA stakeholders are informed about any progress promptly. This thesis could help HTA stakeholders track the progress of addressing RWD-related quality concerns, as we identified all recently developed tools, summarized and compared their signaling questions, and provided suggestions about how these tools could be selected.

In addition, in Chapter 9, we discussed the solutions to the general problems related to HTA method development and implementation. As mentioned in Chapter 1, method developers need approaches that facilitate their understanding on the needs from other HTA stakeholders, i.e., practitioners and beneficiaries. To start with addressing this problem, method developers should acquire insight in limitations of current HTA processes, e.g., by gaining feedback from practitioners who used traditional methods and beneficiaries who are affected by them. Some of the other suggestions, as mentioned in the framework, include picturing what future HTA processes looks like and evaluating heterogeneity of contexts where research gaps are identified. To provide a clear overview of the needs from HTA stakeholders, we also recommend model developers to invite potential method practitioners and future beneficiaries to engage in the process of method development.

To resolve the tension between a long project duration and the relatively insufficient available time of stakeholders, we recommend HTA stakeholders, i.e. developers, practitioners, and beneficiaries, to first list all potential actions needed for developing and implementing a method. To address the limitations due to the lack of some other types of resources (e.g. high-quality data and knowledge), we recommend HTA stakeholders from different HTA contexts to better understand their complementary capabilities of developing or implementing an HTA method, and to collaborate in a larger HTA network.

The lack of expertise or skills to implement a method in the HTA context is a main challenge of HTA method implementation. To tackle this challenge, we recommend developers to inform practitioners and beneficiaries of the contexts (e.g. in terms of geography or therapeutic areas) where using a method is appropriate. To judge in which context implementing a method is appropriate, method developers may first refer to aims, functions, and internal and external validity of a method. Method developers may also share their initial judgement with practitioners and beneficiaries, and check whether their suggestions for the method selection can be adopted or need to be further improved. Finally, a guideline, manual, or interactive tool that guides the method selection can be created, then disseminated to all HTA stakeholders.

In conclusion, the use of the framework developed in this thesis may lead to increased awareness that not only researchers, but all HTA stakeholders (e.g. patients, clinicians and health technology developers) have the potential to contribute to the development or implementation of HTA methods. This framework may also support a more standardized and structured process of innovating HTA methods. In doing so, new challenges in the HTA context may be addressed more efficiently and consistently.

## Samenvatting

Health Technology Assessment (HTA) is een proces waarbij expliciete methoden worden gebruikt om de waarde van een gezondheidstechnologie op verschillende punten in de levenscyclus ervan te bepalen. HTA heeft tot doel bij te dragen aan een efficiënt en rechtvaardig gezondheidszorgsysteem, maar de kwaliteit, relevantie en geschiktheid van HTA-methoden staan vaak ter discussie. Als wetenschappelijk bewijs dat is verkregen en gesynthetiseerd met ongeschikte HTA-methoden wordt gebruikt voor bijvoorbeeld besluitvorming over de vergoeding en prijsstelling van nieuwe gezondheidstechnologieën, zouden belanghebbenden suboptimale beslissingen kunnen nemen. Om de beschikbaarheid van geschikte HTA-methoden te vergroten, zijn HTA-methoden herhaaldelijk (door)ontwikkeld en geïmplementeerd. Dit gebeurt al sinds HTA in de jaren tachtig een belangrijk onderdeel van gezondheidszorgsystemen werd. Om het concept 'HTA-methoden' verder te verduidelijken, hebben we in hoofdstuk 1 een samenvatting gegeven van de manieren waarop 'HTA-methoden' worden genoemd door HTA-organisaties en in onderzoeksprojecten en wetenschappelijke artikelen. Samenvattend omvat het concept de volledige reikwijdte van een HTA-proces, inclusief het verzamelen van wetenschappelijk bewijs, het beoordelen van wetenschappelijk bewijs, besluitvorming en monitoring. Ook kunnen HTA-methoden kwalitatief van aard zijn, denk aan referentiekaders, richtlijnen en checklists, of kwantitatieve modellen en statistische benaderingen omvatten.

Hoewel er in de loop van de tijd veel nieuwe HTA-methoden zijn verschenen, wordt de ontwikkeling en implementatie van deze methoden als suboptimaal ervaren. Eén type probleem dat verband houdt met de ontwikkeling van HTA-methoden is het ontbreken van een duidelijk overzicht van de behoeften van HTA-stakeholders. Hoewel het belang van het begrijpen van de behoeften van HTA-stakeholders steeds meer wordt erkend, ontbreken er nog steeds benaderingen die methode-ontwikkelaars hierbij ondersteunen. Een ander type probleem dat verband houdt met de ontwikkeling van HTA-methoden is de beperking van tijd, beschikbare hoogwaardige gegevens en interdisciplinaire kennis over verschillende onderzoeksgebieden. Bovendien wordt de ontwikkeling van HTA-methoden beperkt door een beperkt inzicht in de theorie hoe HTA-methoden moeten worden ontwikkeld. Gezien de afwezigheid van een alomvattend raamwerk voor het begrip van innovatie-uitdagingen in complex HTA-onderzoek, moet verder onderzoek naar complexe innovatieactiviteiten worden uitgevoerd door HTA-stakeholders. Wat de implementatie van HTA-methoden betreft is een probleem dat gebruikers van de methoden mogelijk geen expertise hebben op het gebied van de HTA-methoden zelf. Bovendien zou de diversiteit aan methoden, die elk specifieke kennis over het gebruik van de methode met zich meebrengen, het probleem

van de implementatie van de methode nog verder kunnen compliceren. Een ander probleem is dat HTA-stakeholders mogelijk niet over samenwerkingsvaardigheden beschikken die voor het gebruik van verschillende onderzoeksmethoden nodig zijn. In het geval van de implementatie van een methode waarbij een grote en diverse groep belanghebbenden betrokken is, is het noodzakelijk de methodologie te verfijnen om tegemoet te komen aan tegenstrijdige standpunten of een mogelijk ontbrekend stakeholderperspectief.

Eén mogelijke oplossing voor de hierboven genoemde algemene problemen is stelselmatig identificeren van de behoeften van de belanghebbenden en het faciliteren van de samenwerking tussen belanghebbenden, gedurende het hele innovatieproces. Eerder onderzoek heeft een aantal fundamenten gelegd door richtlijnen op te stellen voor het ontwikkelen van bepaalde HTA-methoden. Deze onderzoeken hebben echter verschillende beperkingen, zoals het gebrek aan algemene richtlijnen om te begrijpen waarom een methode moet worden ontwikkeld. Daarom is verder onderzoek nodig om algemene richtlijnen te ontwikkelen over hoe HTA-methoden moeten worden geïnnoveerd, waarbij rekening moet worden gehouden met de verschillende typen mehoden. Naast het illustreren van de algemene problemen presenteert Hoofdstuk 1 de behoefte aan methoden om het gebruik van *real-world data* (RWD, klinische praktijkgegevens) in HTA-omgevingen te bevorderen. Omdat RWD een aanvulling kan zijn op gerandomiseerde gecontroleerde onderzoeken (RCT's), vooral wanneer RCT's schaars zijn of niet haalbaar zijn om uit te voeren, is er een groeiende belangstelling voor het gebruik van RWD bij HTA. Het nut van RWD wordt echter vaak in twijfel getrokken vanwege methodologische beperkingen zoals het risico op vertekening. Om het gebruik van RWD te vergemakkelijken, moeten deze kwaliteitsproblemen worden geëvalueerd en vervolgens worden aangepakt met geschikte methoden. Gezien de opkomst van nieuwe gezondheidstechnologieën en de verscheidenheid aan omgevingen waarin HTA wordt uitgevoerd, zoals hierboven vermeld, voldoen de bestaande methoden voor het evalueren of aanpakken van RWD-kwaliteitsproblemen echter mogelijk niet aan alle behoeften van HTA-stakeholders. Daarom moeten er herhaaldelijk nieuwe methoden worden ontwikkeld en geïmplementeerd. Dit proefschrift beoogt deze algemene problemen aan te pakken door relevant conceptueel onderzoek uit te voeren en door te illustreren hoe het conceptuele onderzoek zou kunnen helpen de problemen aan te pakken. Hierbij werd gebruik gemaakt van RWD-gerelateerde HTA-methoden als casus.

Om te beginnen presenteert hoofdstuk 2 een raamwerk met twee functies: het illustreren van een generiek innovatieproces dat toepasbaar is op alle soorten HTA-methoden en het illustreren hoe verschillende HTA-stakeholdergroepen dynamisch

kunnen deelnemen en effectief kunnen samenwerken tijdens het innovatieproces. Het raamwerk is opgebouwd op basis van twintig documenten over vernieuwende HTA-raamwerken en veertien richtlijnen uit drie wetenschappelijke disciplines. Het omvat een generiek innovatieproces dat bestaat uit drie fasen ("Identificatie", "Ontwikkeling", "Implementatie") en negen subfasen. In het raamwerk worden drie rollen gedefinieerd die HTA-stakeholders kunnen spelen bij innovatie ("Ontwikkelaars", "Gebruikers", "Begunstigden") en is een proces opgenomen over hoe de belanghebbenden HTA-methoden innoveren. Dit raamwerk brengt systematisch in beeld welke elementen en stakeholders belangrijk zijn voor de ontwikkeling en implementatie van nieuwe HTA-methoden.

In Hoofdstuk 3 hebben we de validiteit van dit raamwerk verder onderzocht in drie casussen van innovatieve HTA-methoden die deel uitmaken van het HTx-project. De resultaten gaven aan dat het raamwerk een structurele manier van overleg bood en de samenwerking tussen HTA-stakeholders hielp te verbeteren. De toepasbaarheid van het raamwerk zou echter kunnen worden verbeterd, als het wordt aangevuld met een routekaart met een lusstructuur om op maat gemaakte begeleiding te bieden voor verschillende casussen en met extra elementen om acties uit te werken die door belanghebbenden moeten worden genomen. Dienovereenkomstig werd een routekaart met een lusstructuur en achtenveertig bruikbare items ontwikkeld, die het raamwerk zouden kunnen aanvullen en HTA-stakeholders op maat gemaakte begeleiding zouden kunnen bieden bij het ontwikkelen van nieuwe methoden.

In de volgende hoofdstukken hebben we specifieke onderzoeksvragen onderzocht die verband houden met de ontwikkeling of implementatie van kwalitatieve en kwantitatieve HTA-methoden met RWD. Meer specifiek hebben we ons in Hoofdstuk 4,5 en 6 gericht op methoden die worden gebruikt voor het beoordelen van de kwaliteit van onderzoeken met RWD. In Hoofdstuk 7 en 8 hebben we ons gericht op methoden die worden gebruikt voor het samenvoegen van RCT's en RWD in (netwerk)meta-analyses.

In Hoofdstuk 4 hebben we een systematische review uitgevoerd om de methodologische kwaliteit van retrospectieve observationele studies naar de effecten van diabetesmonitoringsystemen te beoordelen en om de trend in kwaliteit van de studies in de loop van de tijd te bestuderen. De resultaten gaven aan dat de algehele methodologische kwaliteit van de observationele studies vrij laag was, aangezien 61 (85%) studies een kritisch of ernstig risico op vertekening hadden. Ook verbeterde de algehele methodologische kwaliteit in de loop van de tijd niet substantieel. De belangrijkste oorzaken van de lage kwaliteit waren onder meer het niet adequaat controleren voor confounding, missende gegevens en selectieve rapportage van de

resultaten. Betrokkenen bij klinische, regelgevende of HTA beslissingen hebben dus mogelijk strategieën nodig om deze suboptimale onderzoeken effectief te benutten. Om de studiekwaliteit te verbeteren kunnen extra inspanningen nodig zijn, zoals ondersteuning bij het gebruik van de juiste tools zodat de kwaliteit van de studies beter kan worden vastgesteld.

In Hoofdstuk 5 hebben we de bestaande beoordelingsinstrumenten voor niet-gerandomiseerde studies van interventies verder geïdentificeerd. Ook hebben we de criteria vergeleken die de instrumenten op het niveau van kwaliteitsitems hanteren. Onze studie identificeerde 49 instrumenten en toonde aan dat de Item Bank van het Research Triangle Institute (RTI) en de STrenngthening the Reporting of OBservational studies in Epidemiology Checklists (STROBE) over het geheel genomen het meest uitgebreid waren, met het hoogste aantal behandelde en voldoende beschreven items ten aanzien van respectievelijk methodologische kwaliteit en rapportage. Geen van de instrumenten richt zich echter op alle kwaliteitsproblemen.. De items die het minst aan bod kwamen voor de methodologische kwaliteit waren uitkomstselectie, uitkomstdefinitie en ethische goedkeuring. Voor de rapportage betrof dit interventieselectie, interventiemeting en duur van de follow-up. De resultaten gaven ook aan dat de meeste beoordelingsinstrumenten hun eigen sterke punten hadden, maar geen van hen kon alle kwaliteitsproblemen adresseren die relevant zijn voor niet-gerandomiseerde studies naar interventies. Zelfs de meest uitgebreide tools kunnen worden aangevuld met andere beschikbare tools. Daarom stellen wij voor dat besluitvormers en onderzoekers bij het selecteren van een hulpmiddel of tool rekening houden met de heterogeniteit op het niveau van kwaliteitsitems.

In Hoofdstuk 6 hebben we een literatuuronderzoek uitgevoerd om modellen te identificeren voor het voorspellen van het risico op coronaire hartziekten bij patiënten met diabetes. Vervolgens hebben we de kwaliteit van de modellen beoordeeld in termen van risico op vertekening en bruikbaarheid voor HTA. Van de 26 modelstudies en 30 geïdentificeerde modellen had slechts één modelstudie een laag risico op vertekening in alle domeinen en geen enkel model was volledig bruikbaar voor HTA. Wij adviseren om bij het ontwikkelen van toekomstige voorspellingsmodellen rekening te houden met de behoeften van HTA-stakeholders, in het bijzonder in relatie tot gebruik in gezondheidseconomische modellen en na validatie met bestaande kwaliteitsbeoordelingsinstrumenten. Omdat deze voorspellingsmodellen in het algemeen nog onvoldoende bruikbaar zijnvoor HTA, moeten er waarschijnlijk nieuwe en/of aangepaste instrumenten worden ontwikkeld.

In Hoofdstuk 7 hebben we parallelle netwerk-meta-analyses uitgevoerd naar de effectiviteit van diabetesmonitoringsystemen met insulinetoediening bij patiënten met diabetes mellitus type 1, waarbij we niet-gerandomiseerde onderzoeken (NRS's), RCT's of de combinatie van deze twee als wetenschappelijk bewijs gebruikten. We onderzochten of de geschatte effectiviteit veranderde op basis van de gebruikte data in de meta-analyse. Uit het onderzoek bleek dat RCT's in twee afzonderlijke netwerken werden geanalyseerd en dat ze pas na de integratie met de NRS's in één netwerk konden worden opgenomen. Nadat het relatieve gewicht van de NRS's was verlaagd in de meta-analyse, konden de NRS's de interventienetwerken van RCT's samenvoegen en uitbreiden. Bovendien verschilden de geschatte effectiviteit en rangschikking van effectiviteit uit NRSs en RCTs, maar waren deze verschillen niet statistisch significant. Daarentegen waren de resultaten van RCTs en gecombineerd wetenschappelijk bewijs grotendeels vergelijkbaar. Tenslotte had het veranderen van het gewicht van NRSs ten opzichte van RCT's, vooral voor die NRSs met een ernstig risico op bias, een grote en statistisch significante invloed op de geschatte effectiviteit..

In Hoofdstuk 8 hebben we een R-pakket (dwz Crossnma) toegepast op casussen, te weten voor myelodysplastische syndromen en diabetes(2 per ziektegebied, in het totaal 4). , om te onderzoeken of de effectschattingen verkregen met drie statistische benaderingen die de synthese van RCT's en NRS's ondersteunen (naïeve pooling, power prior en hiërarchische modellering) tot consistente resultaten leiden. Het onderzoek toonde aan dat de 'power prior'-benadering betrouwbaarder was dan de naïeve benadering van pooling en hiërarchische modellering voor het synthetiseren van wetenschappelijk bewijs uit RCT's en NRS's in een meta-analyse, maar geen van de benaderingen kon de nauwkeurigheid van gepoolde schattingen garanderen wanneer het absolute aantal of de proportie van de RCT's klein was.

Gebaseerd op de bevindingen uit voorgaande hoofdstukken hebben we in Hoofdstuk 9 de implicaties besproken die verband houden met het gebruik van RWD in de HTA-context. Onze bevindingen uit hoofdstuk 4-8 kunnen het gebruik van RWD vooral op twee aspecten helpen verbeteren. Ten eerste zouden de resultaten kunnen bijdragen aan de ontwikkeling van beleid voor het gebruik van RWD bij HTA. Eén doel van het beleid van de HTA en regelgevende instanties is om HTA-stakeholders te helpen bij het aanpakken van methodologische uitdagingen en de selectie van een geschikte methode om HTA. Ons onderzoek naar de innovatie van instrumenten om de kwaliteit van studies met RWD te beoordelen ėn naar benaderingen om gegevens uit RCT's en RWD samen te voegen, zou als startpunt kunnen dienen voor het actualiseren van het huidige HTA-beleid inzake het verzamelen van wetenschappelijk bewijs. Hiertoe verwijzen we naar het overzicht dat we gaven van de bestaande methoden (hoofdstuk 5 en 8)

en de illustratie hoe ze gebruikt konden worden (hoofdstuk 4, 5 en 7). Op vergelijkbare wijze zou het onderzoek dat in dit proefschrift wordt gepresenteerd input kunnen leveren voor het beleid ten aanzien van gezondheidseconomische analyses Dit betreft ons onderzoek naar risicovoorspellingsmodellen (hoofdstuk 6) en hoe een model dat risico kan voorspellen methodologisch verantwoord kan worden geïncorporeerd in gezondheidseconomische modellen. Ten tweede kunnen de HTA-methoden die in ons onderzoek worden beschreven helpen bij het aanpakken van enkele uitdagingen bij het gebruik van RWD ten behoeve van HTA. Een uitdaging bij het gebruik van RWD in de HTA-context is bijvoorbeeld de moeilijkheid om kwaliteitsproblemen met betrekking tot RWD te evalueren. Deze uitdaging kan geleidelijk worden overwonnen als belanghebbenden zowel binnen als buiten de HTA-context voortdurend inspanningen leveren voor de ontwikkeling, selectie en implementatie van robuuste kwaliteitsbeoordelingsinstrumenten (bijvoorbeeld door gebruik te maken van RWD voor regelgevingsdoeleinden). Tegelijkertijd zouden HTA-stakeholders tijdig moeten worden geïnformeerd over eventuele vorderingen. Dit proefschrift zou HTA-stakeholders tot slot ook kunnen helpen de voortgang bij het aanpakken van RWD-gerelateerde kwaliteitsproblemen te volgen, aangezien we alle recent ontwikkelde instrumenten hebben geïdentificeerd, hun signaalvragen hebben samengevat en vergeleken en suggesties hebben gegeven over hoe deze instrumenten kunnen worden geselecteerd.

Daarnaast hebben we in Hoofdstuk 9 de oplossingen besproken voor de algemene problemen die verband houden met de ontwikkeling en implementatie van HTA-methoden. Zoals vermeld in Hoofdstuk 1 hebben methodeontwikkelaars een aanpak nodig die hun inzicht in de behoeften van andere HTA-stakeholders, dat wil zeggen gebruikers en begunstigden, vergemakkelijkt. Om te beginnen met het aanpakken van dit probleem moeten methodeontwikkelaars inzicht verwerven in de beperkingen van de huidige HTA-processen, bijvoorbeeld door feedback te krijgen van gebruikers van traditionele HTA methoden en van degenen die de resultaten van deze analyses voor beleid gebruiken. Enkele van de andere suggesties, zoals genoemd in het raamwerk, omvatten het in kaart brengen van hoe toekomstige HTA-processen eruit kunnen zien en het evalueren van de heterogeniteit van contexten waarin lacunes in het onderzoek worden geïdentificeerd. Om een duidelijk overzicht te geven van de behoeften van HTA-stakeholders raden we modelontwikkelaars ook aan om potentiële gebruikers en toekomstige begunstigden uit te nodigen om deel te nemen aan het proces van methodeontwikkeling.

Om de spanning tussen een lange projectduur en de relatief beperkte beschikbare tijd van belanghebbenden op te lossen, raden wij HTA-stakeholders, dat wil zeggen

ontwikkelaars, gebruikers en begunstigden, aan om eerst alle potentiële acties op te sommen die nodig zijn voor het ontwikkelen en implementeren van een methode. Om de beperkingen aan te pakken die het gevolg zijn van het ontbreken van een aantal andere benodigdheden (bijvoorbeeld gegevens van hoge kwaliteit en hoogwaardige kennis), raden we HTA-stakeholders uit verschillende HTA-contexten aan om hun complementaire capaciteiten bij het ontwikkelen of implementeren van een HTA-methode beter te begrijpen en om samen te werken een groter HTA-netwerk.

Het gebrek aan expertise of vaardigheden om een methode in de HTA-context te implementeren is een van de belangrijkste uitdagingen bij de implementatie van HTA-methoden. Om deze uitdaging aan te pakken raden we ontwikkelaars aan om gebruikers en begunstigden te informeren over de context (bijvoorbeeld in termen van geografie of therapeutische gebieden) waarvoor de inzet van een methode relevant is. Om te beoordelen in welke context het implementeren van een methode passend is, kunnen methodeontwikkelaars eerst verwijzen naar de doelstellingen, functies en interne en externe validiteit van een methode. Methodeontwikkelaars kunnen ook hun aanvankelijke oordeel delen met gebruikers en begunstigden en nagaan of hun suggesties voor de methodeselectie kunnen worden overgenomen of verder verbeterd moeten worden. Ten slotte kan er een richtlijn, handboek of interactief hulpmiddel worden ontwikkeld dat de methodekeuze begeleidt en dat vervolgens onder alle HTA-belanghebbenden wordt verspreid.

Concluderend kan het gebruik van het in dit proefschrift ontwikkelde raamwerk leiden tot een groter besef dat niet alleen onderzoekers, maar alle HTA-stakeholders (bijv. patiënten, artsen en ontwikkelaars van gezondheidstechnologie) het potentieel hebben om bij te dragen aan de ontwikkeling of implementatie van HTA-methoden. Dit raamwerk kan ook een meer gestandaardiseerd en gestructureerd proces van innovatieve HTA-methoden ondersteunen. Door dit te doen kunnen nieuwe uitdagingen in de HTA-context efficiënter en consistenter worden aangepakt.

# Acknowledgements

I would like to thank Wim, Aukje, Olaf, Junfeng, and Rick for providing constant and valuable guidance to my research during my PhD.

I would like to thank Milou, Jan-Willem, and Marcelien for contributing to my research and providing valuable inputs.

I would like to thank my colleagues at Universidad Politécnica de Madrid (UPM), University of York (UoY), Medical University of Sofia (MUS), National Institute of Health and Care Excellence (NICE), Syreon Research Institute (SRI), National Health Care Institute (ZIN), and European Organisation for Research and Treatment of Cancer (EORTC), for their contribution to my research.

I would like to thank my parents and my wife for constantly providing spiritual support during my PhD.

# List of publications

## Scientific articles

Jiu L, Hogervorst MA, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. Understanding innovation of health technology assessment methods: the IHTAM framework. International Journal of Technology Assessment in Health Care. 2022;38(1):e16.

Jiu L, Hartog M, Wang J, Vreman RA, Klungel OH, Mantel-Teeuwisse AK, Goettsch WG. Tools for assessing quality of studies investigating health interventions using real-world data: a literature review and content analysis. BMJ open. 2024;14(2):e075173.

Jiu L, Wang J, Somolinos-Simón FJ, Tapia-Galisteo J, García-Sáez G, Hernando M, Li X, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. A literature review of quality assessment and applicability to HTA of risk prediction models of coronary heart disease in patients with diabetes. Diabetes Research and Clinical Practice. 2024:111574.

Jiu L, Wang J, Kamusheva M, Dimitrova M, Tachkov K, Milushewa P, Mitkova Z, Petrova G, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. Methodological Quality of Retrospective Observational Studies Investigating Effects of Diabetes Monitoring Systems: a Systematic Review. 2024 Feb * (under review).

Jiu L, Wang J, Versteeg J, Jin J, Deng Y, Tashkov K, Petrova G, Klungel OH, Mantel-Teeuwisse AK, Goettsch WG. Comparison of Network Meta-analyses Investigating Efficacy of Diabetes Monitoring Systems with Insulin Delivery in Patients with Type-1 Diabetes, using Non-randomized Studies, Randomized-controlled Trials, or Both as Evidence. 2024 Feb * (under review).

Jiu L, Wang J, Versteeg J, Zhang Y, Liu L, Somolinos-Simón FJ, Tapia-Galisteo J, García-Sáez G, Hogervorst MA, Li X, Mantel-Teeuwisse AK, Goettsch WG. Roadmap to Innovation of HTA Methods (IHTAM): Insights from Three Case Studies of Quantitative Methods. 2024 Feb * (under review).

Jiu L, Wang J, Klungel OH, Mantel-Teeuwisse AK, Goettsch WG. Approaches to Synthesizing Evidence from Randomized Controlled Trials and Non-randomized Studies in Meta-analyses: Application of the Crossnma Package to the Cases of Myelodysplastic Syndromes and Diabetes. 2024 Feb * (article prepared for submission).

**Project report**

Goettsch WG, Mantel-Teeuwisse AK, Hogervorst MA, Vreman RA, Jiu L. WP1: Deliverable D1.2. General framework definition for case studies. Utrecht (NL): HTx; 2020. [cited 2024 Feb 5].

# List of co-authors

(The list is presented in alphabetical order and affiliations listed are those at the time that the studies were conducted)

**Yingnan Deng**
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

**Maria Dimitrova**
Faculty of Pharmacy, Medical University of Sofia, Sofia, Bulgaria

**Gema García-Sáez**
Bioengineering and Telemedicine Group, Centro de Tecnología Biomédica, ETSI de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain CIBER-BBN: Networking Research Centre for Bioengineering, Biomaterials and Nanomedicine, Madrid, Spain

**Wim G Goettsch**
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht; The National Health Care Institute, Diemen, The Netherlands

**Michiel Hartog**
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

**Mariaelena Hernando**
Bioengineering and Telemedicine Group, Centro de Tecnología Biomédica, ETSI de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain CIBER-BBN: Networking Research Centre for Bioengineering, Biomaterials and Nanomedicine, Madrid, Spain

**Milou A Hogervorst**
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

**Jing Jin**
Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, The Netherlands

**Maria Kamusheva**
Faculty of Pharmacy, Medical University of Sofia, Sofia, Bulgaria

**Olaf H Klungel**
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

**Xinyu Li**
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands University of Groningen, Faculty of Science and Engineering, Groningen Research Institute of Pharmacy, Groningen, The Netherlands

**Lifang Liu**
The European Organisation for Research and Treatment of Cancer, Brussels, Belgium

**Aukje K Mantel-Teeuwisse**
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

**Petya Milushewa**
Faculty of Pharmacy, Medical University of Sofia, Sofia, Bulgaria

**Zornitsa Mitkova**
Faculty of Pharmacy, Medical University of Sofia, Sofia, Bulgaria

**Guenka Petrova**
Faculty of Pharmacy, Medical University of Sofia, Sofia, Bulgaria

**Francisco Javier Somolinos-Simón**
Bioengineering and Telemedicine Group, Centro de Tecnología Biomédica, ETSI de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain

**Konstantin Tachkov**
Faculty of Pharmacy, Medical University of Sofia, Sofia, Bulgaria

**Jose Tapia-Galisteo**
Bioengineering and Telemedicine Group, Centro de Tecnología Biomédica, ETSI de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain CIBER-BBN: Networking Research Centre for Bioengineering, Biomaterials and Nanomedicine, Madrid, Spain

**Jan-Willem Versteeg**
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

**Rick A Vreman**
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands National Health Care Institute, Diemen, The Netherlands

**Junfeng Wang**
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

**Yingying Zhang**
Centre for Health Economics, University of York, York, United Kingdom

# About the author

Li Jiu was born in Nanjing, China. He completed his bachelor's studies in International Economics and Trade at the China Pharmaceutical University in Nanjing in 2016. Afterwards, Li moved to Rotterdam, the Netherlands, where he completed his master's studies in Health Sciences at Erasmus University Rotterdam in 2019. Throughout his master's studies, Li has conducted research in the field of early health technology assessment. In December 2019, he started his PhD trajectory at UIPS' division of Pharmacoepidemiology and Clinical Pharmacology, Faculty of Science, Utrecht University. Li conducted his studies mainly in three subfields: conceptual research on methods of health technology assessment, quality assessment tools for non-randomized studies, and application of risk prediction models to a health economics model. The research was performed as part of the HTx project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825162.

Next to his PhD studies, Li often reads history books and enjoys Spanish art.