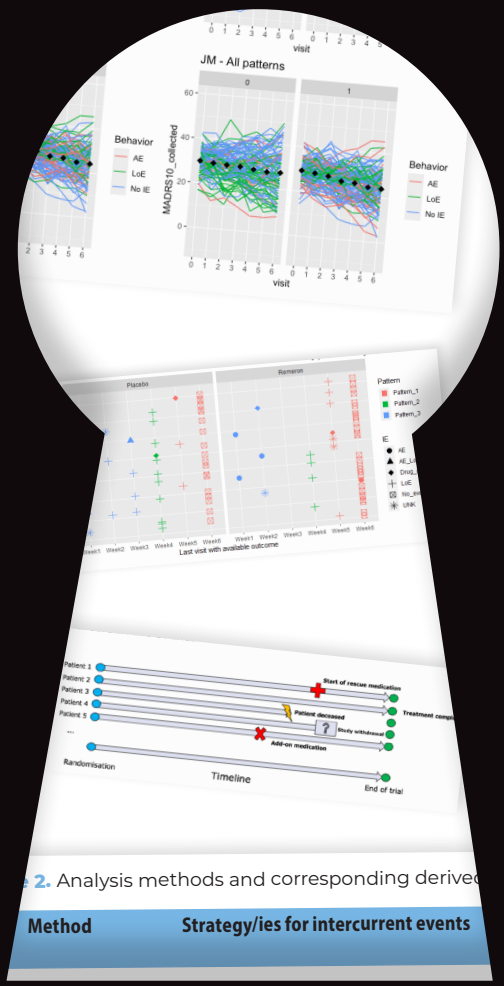
Estimands in clinical drug development: from design to regulatory assessment



Marian Mitroiu

Estimands in clinical drug development: from design to regulatory assessment

Marian Mitroiu

Estimands in clinical drug development: from design to regulatory assessment
PhD Dissertation, Utrecht University, The Netherlands

ISBN: 978-90-393-7528-0
DOI: https://doi.org/10.33540/723
Author: Marian Mitroiu
Printing: Gildeprint

The research described in this thesis was supported by Dutch Medicines Evaluation Board (CBG-MEB).

Financial support by the Julius Center, UMC Utrecht for the publication of this thesis is gratefully acknowledged.

Estimands in clinical drug development: from design to regulatory assessment

Estimands in klinische geneesmiddelenontwikkeling: van design tot regulatoire beoordeling
(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

maandag 12 december 2022 des ochtends te 10.15 uur

door

Marian Mitroiu

geboren op 18 mei 1988
te Buzău, Roemenië

Promotor:
Prof. dr. K.C.B. Roes

Copromotoren:
Dr. K. Oude Rengerink
Dr. S. Teerenstra

Beoordelingscommissie:
Prof. dr. A.C.G. Egberts
Prof. dr. O.H. Klungel
Prof. dr. D. Rizopoulos
Prof. dr. R.H.H. Groenwold
Dr. N. Benda

Contents

subitise (ˈsʌbɪˌtaɪz)

verb

to perceive the number of presented items at a glance and without counting

CHAPTER 1

General Introduction

Estimands – from the first randomised controlled clinical trial to present times

Randomised clinical trials are the most widely used experiments to investigate efficacy and safety of innovative or repurposed treatments [1]. They are commonly used as basis for regulatory evaluation and approval of medicinal products, due to their advantages over other types of designs [2].

In 1948 the "*short-term results of a controlled investigation into the effects of streptomycin on one type of pulmonary tuberculosis*" were reported in the British Medical Journal (BMJ). This represents a methodological landmark in clinical research, and it represents the cornerstone principles of the statistical framework for clinical trials [3]. It is the first randomised controlled clinical trial reported in the scientific literature. It compared streptomycin additional to bed rest with bed rest alone, in patients suffering from pulmonary tuberculosis. The main reasons to conduct a controlled trial was that the outcome of tuberculosis infection was deemed heterogeneous and unpredictable, and because of previously conducted trials that had methodological flaws [4].

It used inclusion and exclusion criteria for patients' screening and recruitment, with the main aim "*to eliminate as many of the obvious variations as possible*". The treatment assignment "*was made by reference to a statistical series based on random sampling numbers drawn up for each sex at each centre by Professor Bradford Hill*". It used blinding, patients were not informed before admission that they were undergoing the investigation and patients allocated to different treatments were not usually kept in the same ward. Doctors were not blinded and were specifically asked to keep all information confidential. Patients were to remain admitted at the centre and followed-up for at least six months, with their clinical outcome assessed after this period. Each patient had what would constitute today a case report form, where each patient's characteristics were recorded, from medical history to adverse events and clinical outcome. Data were monitored constantly, and monthly meetings were conducted to discuss the progress of the trial. Independent assessors of the blood samples were also engaged in this investigation. For the statistical analysis, although not reported in the published BMJ article, it appears that Professor Bradford Hill might have used a chi-square test to analyse the trial results and estimate the treatment effect [5].

Randomisation is one of the key principles used in the MRC 1948 streptomycin trial. It ensures the treatment assignment is independent of patients' characteristics, which may be (known or unknown) prognostic variables for the clinical outcome, such that estimated treatment effect would be attributable solely to the treatment [6].

In any trial and for various known or unknown, expected or unexpected reasons, patients may withdraw from the study, or stay in the study but not adhere to their treatments as formalised in the protocol. They may discontinue their assigned treatments, take rescue or prohibited medication, or simply switch to other available treatments. All these unplanned changes or events were commonly known as "post-randomisation events" and can influence the estimation and clinical meaning of the treatment effect [7–9].

One argument to use the treatment effect estimate regardless of post-randomisation events is that they will also be experienced by patients (after the medicinal product is approved and used) in usual clinical practice [10]. It is however often unknown to what extent this will also be expected in practice, once the medicinal product is actually used at large. This estimated treatment effect is known as "*Intention-To-Treat principle*" (ITT), a term that was described formally in the context of late-stage drug development by the ICH E9 guideline Statistical Principles for Clinical Trials [8]. The ITT effect takes into account patients' outcomes irrespective of their behavior in the trial or adherence to the assigned treatment [11–13]. Even if patients were present at all visits, some of them may experience adverse events and would not be able to continue taking the drug, while others may not perceive any benefit from taking it and discontinue the treatment. The ITT respects the randomisation principle and even though this was acceptable for the scientific community for a long time, it is no longer seen as the only treatment effect of interest in a clinical trial [14–16]. It needed to change and evolve, situation that led to the development and introduction of the estimands framework as an addendum to ICH E9 guideline [17].

In 2011, regulatory evaluation of a dossier for an antidiabetic drug included Phase III trials in which patients that took rescue medication while being in the trial were handled by the sponsor differently than the regulators preferred. The sponsor excluded the patients' outcome data following rescue medication intake and imputed the remainder of the resulting missing outcomes with last observation carried forward (LOCF). However, the regulators (US FDA) wanted to use the collected outcome data as it was after rescue medication intake, without "creating" and imputing missing data [8]. While both appear to respect the randomisation principle (analysing patients as they were randomised), there is a more subtle and fundamental difference between the two approaches. The sponsor is interested in the treatment effect "if the patients were not to take rescue medication while under treatment", and the regulator is interested in the treatment effect "when patients took rescue medication as needed additional to the treatment". This brings to light that different stakeholders can be interested in different treatment effects corresponding to different clinical questions without being aware a priori, while trying to preserve the randomisation principle. In absence of a common understanding

and a clear and mutually agreed scientific question, it is to be expected that confusion can occur [18].

Without any doubt, it is often of interest to know the ITT effect [10,13,19,20]. However, it may also be relevant to know, for instance, what would be the treatment effect in those patients that are able to go through a certain toxicity period of a treatment where they do not have any benefit, what would be the treatment effect if they were not to (need to) take rescue medication, or the treatment effect prior to treatment discontinuation [21–24].

In 2014, the ICH E9(R1) expert working group published the concept paper on the need for an addendum to ICH E9 guideline, namely for the need of an estimands framework [25]. In 2017, the draft Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials E9(R1) was released for public consultation and specified the estimand as being described by four attributes: Variable, intercurrent event and strategies for intercurrent events, population and population-level summary [26]. In late 2019, the final version - Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials E9(R1) was published and started to become effective in the regulatory world the following year [17,27]. The estimand is defined as *"A precise description of the treatment effect reflecting the clinical question posed by the trial objective. It summarizes at a population-level what the outcomes would be in the same patients under different treatment conditions being compared."* It is described by five attributes: the "treatment" was added in the final version to describe the treatment plan for patients according to the trial design and objective; the "population" represents the target population of patients; "variable" is the measurement of the clinical outcome, "population-level summary" is the numerical estimate of the treatment effect. The "strategies for intercurrent events" are used to handle the intercurrent events when describing the clinical question of interest. The addendum suggests five possible strategies: treatment policy, hypothetical, composite variable, (generic) while on treatment and principal stratum. The treatment policy actively ignores the intercurrent event, hypothetical envisages a scenario where the intercurrent event would not occur, composite variable takes the intercurrent event into account at the level of the variable, while on treatment uses all clinical outcomes prior to the intercurrent event and principal stratum refers to a population of interest that would not experience a certain intercurrent event, such as lack of efficacy.

The addendum defines the intercurrent events as "Events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest." (Figure 1). And it states a key activity

crucial for any trial: "It is necessary to address intercurrent events when describing the clinical question of interest in order to precisely define the treatment effect that is to be estimated."

The estimands framework was developed to add clarity to what treatment effects are actually being investigated and estimated from the data generated by a clinical trial. Even though the E9(R1) was not available when the MRC 1948 streptomycin trial was conducted, the trial already included some elements of the estimand framework. Treatment was "patients would be treated by streptomycin and bed-rest or by bed-rest alone". Population was "patients suffering from pulmonary tuberculosis (defined by inclusion/exclusion criteria at trial level)". Variable was "radiological assessment of improvement/deterioration or death". Population-level summary was (likely) "difference in counts of improvement/deterioration and deaths between streptomycin and control". It is not possible to derive strategies for intercurrent events based on available information.

Even though the clinical trials have been a cornerstone of clinical research for many decades, and some elements of the estimand have been present all along, there are still challenges to be addressed for a proper estimation of treatment effects: intercurrent events and strategies to handle them [28,29].

For a multidisciplinary discussion on the question the trial seeks to address, the estimands framework promises to be an important instrument in planning, designing, conducting trials, analysing the collected data and interpreting the estimated treatment effect [30–37].

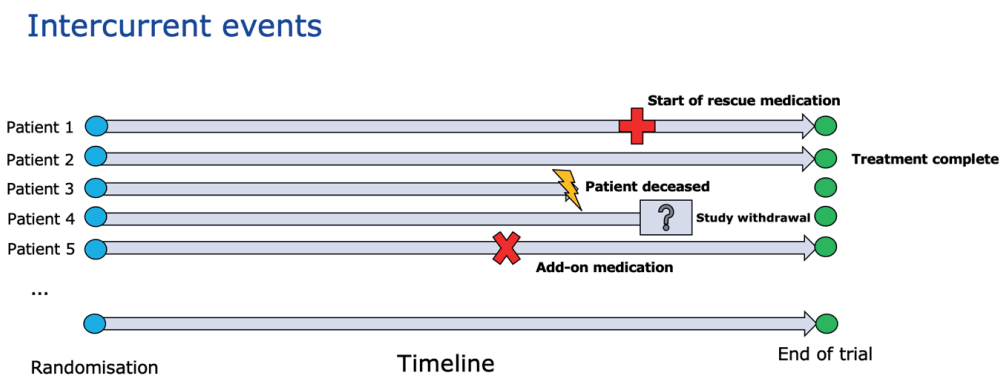


Figure 1. Patient trajectories with intercurrent events

Objectives

In this thesis we investigate and evaluate the estimands practices, we identify, formulate, evaluate different estimands and formulate corresponding clinical questions, and we investigate methods to compare different estimands. We aim to contribute to the implementation of the estimands methodology in clinical trials for regulatory purposes. The main objectives of the thesis included:

- Evaluate established regulatory practice of treatment effect estimation against the estimand framework, by investigating disease guidelines, sponsor marketing authorisation documentation and regulatory questions used in drug evaluation and approval;
- Understand what estimands, if any, correspond to commonly employed analysis methods used for primary efficacy analyses, using trial data from anti-depression trials as example;
- Develop different data-generating models (DGM) that simultaneously generate outcomes and intercurrent events, to facilitate simulations studies for estimands;
- Provide some recommendations to facilitate understanding and implementation of the estimands framework in drug development, regulatory evaluation and approval.

Thesis Outline

In **Chapter 2**, we constructively critique the ITT and per protocol analyses, and promote broadly the estimands framework as a solution for other treatment effects of interest – other estimands, beyond ITT especially [14].

In **Chapter 3**, in our narrative review we aim to understand what kind of estimands, especially strategies for intercurrent events, are used in European Medicines Agency (EMA) disease guidelines, used by pharmaceutical companies in their drug development programs and additionally requested by EMA during centralized procedure regulatory assessment in the so-called "Day 120 List of Questions" (regulatory questions).

With this research we review the current state of affairs in treatment effect estimation practices in drug development and regulatory approval before the final addendum was published [38].

In **Chapter 4**, we aim to understand what estimands correspond to common efficacy analyses employed in clinical trials to estimate treatment effects for regulatory evaluation and approval purposes. Trials for an anti-depression treatment are used as leading example. This research aims to clarify whether any estimands correspond to the most common efficacy analyses employed in trials, which are these corresponding estimands, and to understand the treatment effects and assumptions behind each analysis method [39].

In **Chapter 5**, we develop data-generating models (DGMs) to simulate randomised clinical trials and model the association between efficacy outcomes and intercurrent events. This research is needed in order to enable analytical or simulation studies that can evaluate and compare estimand strategies on datasets generated under the same data generating models.

With our data-generating models and simulation studies we provide the tools to be used in planning and designing studies that can aim at guideline-recommended estimands or at different estimands and that can help identify the best estimator for the formulated and targeted estimand. Furthermore, it can help formulate meaningful sensitivity analyses that target the same estimand or possibly different meaningful supplementary analyses that target different estimands.

References

1. Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? BMJ. 1999 May 1;318(7192):1209.
2. Temple R, Ellenberg SS. Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments. Part 1: Ethical and Scientific Issues. Ann Intern Med. 2000 Sep 19;133(6):455–63.
3. Hill AB (1937) [Internet]. The James Lind Library. 2010 [cited 2021 Jun 24]. Available from: https://www.jameslindlibrary.org/hill-ab-1937b/
4. Streptomycin Treatment of Pulmonary Tuberculosis. Br Med J. 1948 Oct 30;2(4582):769–82.
5. A statistical note on the analysis of the 1948 MRC streptomycin trial. [Internet]. The James Lind Library. [cited 2021 Jun 23]. Available from: https://www.jameslindlibrary.org/articles/a-note-on-the-statistical-analysis-of-the-1948-mrc-streptomycin-trial/
6. Altman DG. Randomisation. BMJ. 1991 Jun 22;302(6791):1481–2.
7. Campbell RT, Willox GP, Jhund PS, Hawkins NM, Huang F, Petrie MC, et al. Reporting of Lost to Follow-Up and Treatment Discontinuation in Pharmacotherapy and Device Trials in Chronic Heart Failure: A Systematic Review. Circ Heart Fail [Internet]. 2016 May [cited 2021 Jun 23];9(5). Available from: https://www.ahajournals.org/doi/10.1161/CIRCHEARTFAILURE.115.002842
8. Akacha M, Bretz F, Ruberg S. Estimands in clinical trials - broadening the perspective. Statistics in Medicine. 2017 Jan 15;36(1):5–19.
9. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. The Lancet. 2002 Mar;359(9308):781–5.
10. Lewis JA, Machin D. Intention to treat--who should use ITT? Br J Cancer. 1993 Oct;68(4):647–50.
11. Leuchs A-K, Brandt A, Zinserling J, Benda N. Disentangling estimands and the intention-to-treat principle. Pharmaceutical Statistics. 2017 Jan 1;16(1):12–9.
12. Gravel J, Opatrny L, Shapiro S. The intention-to-treat approach in randomized controlled trials: are authors saying what they do and doing what they say? Clin Trials. 2007;4(4):350–6.
13. Ten Have TR, Normand S-LT, Marcus SM, Brown CH, Lavori P, Duan N. Intent-to-Treat vs. Non-Intent-to-Treat Analyses under Treatment Non-Adherence in Mental Health Randomized Trials. Psychiatr Ann. 2008 Dec;38(12):772–83.
14. Oude Rengerink K, Mitroiu M, Teerenstra S, Pétavy F, Roes KCB. Rethinking the intention-to-treat principle: one size does not fit all. Journal of Clinical Epidemiology. 2020 Sep;125:198–200.
15. Pétavy F, Guizzaro L, Reis IA dos, Teerenstra S, Roes KCB. Beyond "Intent-to-treat" and "Per protocol": Improving assessment of treatment effects in clinical trials through the specification of an estimand. British Journal of Clinical Pharmacology. 2020;86(7):1235–9.
16. Day S. Editorial: Estimands in clinical trials – a challenge to intention to treat? Statistics in Medicine. 2017;36(1):3–4.
17. ICH E9(R1) EWG. ADDENDUM ON ESTIMANDS AND SENSITIVITY ANALYSIS IN CLINICAL TRIALS TO THE GUIDELINE ON STATISTICAL PRINCIPLES FOR CLINICAL TRIALS E9(R1) [Internet]. Available from: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf
18. Mehrotra DV, Hemmings RJ, Russek-Cohen E. Seeking harmony: estimands and sensitivity analyses for confirmatory clinical trials. Clinical Trials. 2016 Aug 1;13(4):456–8.
19. Glossary [Internet]. NICE. NICE; [cited 2021 Jun 23]. Available from: https://www.nice.org.uk/glossary?letter=i
20. Gross D, Fogg L. A Critical Analysis of the Intent-to-Treat Principle in Prevention Research. The Journal of Primary Prevention. 2004;25(4):475–89.
21. Bornkamp B, Rufibach K, Lin J, Liu Y, Mehrotra DV, Roychoudhury S, et al. Principal stratum strategy: Potential role in drug development. Pharmaceutical Statistics. 2021 Feb 23;pst.2104.
22. Larsen KG, Josiassen MK. A New Principal Stratum Estimand Investigating the Treatment Effect in Patients Who Would Comply, If Treated With a Specific Treatment. Statistics in Biopharmaceutical Research. 2020 Jan 2;12(1):29–38.
23. Darken P, Nyberg J, Ballal S, Wright D. The attributable estimand: A new approach to account for intercurrent events. Pharmaceutical Statistics [Internet]. 2020 Mar 21 [cited 2020 Mar 23]; Available from: http://doi.wiley.com/10.1002/pst.2019

24. Qu Y, Luo J, Ruberg SJ. Implementation of tripartite estimands using adherence causal estimators under the causal inference framework. Pharmaceutical Statistics. 2021;20(1):55–67.

25. ICH E9(R1) EWG. Final Concept Paper E9(R1): Addendum to Statistical Principles for Clinical Trials on Choosing Appropriate Estimands and Defining Sensitivity Analyses in Clinical Trials, Oct. 2014 [Internet]. 2014 [cited 2018 Jan 17]. Available from: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9__R1__Final_Concept_Paper_October_23_2014.pdf

26. ICH E9(R1) EWG. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Step 2b [Internet]. [cited 2017 Sep 8]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/08/WC500233916.pdf

27. ICH Official web site : ICH [Internet]. [cited 2021 May 17]. Available from: https://www.ich.org/page/efficacy-guidelines

28. Qu Y, Shurzinske L, Sethuraman S. Defining estimands using a mix of strategies to handle intercurrent events in clinical trials. Pharmaceutical Statistics [Internet]. [cited 2020 Oct 25];n/a(n/a). Available from: http://onlinelibrary.wiley.com/doi/abs/10.1002/pst.2078

29. Qu Y, Lipkovich I. Implementation of ICH E9 (R1): A Few Points Learned During the COVID-19 Pandemic. Ther Innov Regul Sci [Internet]. 2021 May 13 [cited 2021 May 14]; Available from: https://doi.org/10.1007/s43441-021-00297-6

30. Phillips A, Abellan-Andres J, Soren A, Bretz F, Fletcher C, France L, et al. Estimands: discussion points from the PSI estimands and sensitivity expert group. Pharm Stat. 2017 Jan;16(1):6–11.

31. Lawrance R, Degtyarev E, Griffiths P, Trask P, Lau H, D'Alessio D, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? J Patient Rep Outcomes. 2020 Dec;4(1):68.

32. Keene ON, Ruberg S, Schacht A, Akacha M, Lawrance R, Berglind A, et al. What matters most? Different stakeholder perspectives on estimands for an invented case study in COPD. Pharm Stat. 2020 Jan 9;

33. Callegari F, Akacha M, Quarg P, Pandhi S, von Raison F, Zuber E. Estimands in a Chronic Pain Trial: Challenges and Opportunities. Statistics in Biopharmaceutical Research. 2019 Jul 22;1–6.

34. Akacha M. Choosing Measures of Treatment Benefit: Estimands and Beyond. CHANCE. 2019 Oct 2;32(4):12–7.

35. Fiero MH, Pe M, Weinstock C, King-Kallimanis BL, Komo S, Klepin HD, et al. Demystifying the estimand framework: a case study using patient-reported outcomes in oncology. The Lancet Oncology. 2020 Oct;21(10):e488–94.

36. Rosenkranz G. Estimands—new statistical principle or the emperor's new clothes? Pharmaceutical Statistics. 2017 Jan 1;16(1):4–5.

37. Scharfstein DO. A constructive critique of the draft ICH E9 Addendum. Clinical Trials. 2019 Jun 11;1740774519853566.

38. Mitroiu M, Oude Rengerink K, Teerenstra S, Pétavy F, Roes KCB. A narrative review of estimands in drug development and regulatory evaluation: old wine in new barrels? Trials. 2020 Dec;21(1):671.

39. Mitroiu M, Teerenstra S, Oude Rengerink K, Pétavy F, Roes KCB. Estimation of treatment effects in short-term depression studies. An evaluation based on the ICH E9(R1) estimands framework. Pharmaceutical Statistics. 2022;1-21. doi:10.1002/pst.2214

CHAPTER 2

Rethinking the intention-to-treat principle: one size does not fit all

Oude Rengerink K, Mitroiu M, Teerenstra S, Pétavy F, Roes KCB. Rethinking the intention-to-treat principle: one size does not fit all. Journal of Clinical Epidemiology. 2020 Sep;125:198–200. https://doi.org/10.1016/j.jclinepi.2020.04.023

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) released ''Estimands and sensitivity analyses'', as an addendum to their ICH E9 guideline ''Statistical Principles for Clinical Trials'' [1]. Importantly, we have recently seen the estimand entering clinical trial reports in medical journals, an example shown in Box 1 [2-5]. Although the concept of estimands is not new [6], its adoption by regulatory authorities and the pharmaceutical industry is expected to lead to global changes in trial design, trial conduct, and interpretation of results. As these changes will not only affect trialists but also clinicians, patients, and others using trial results in their decision making, this change is too big to go unnoticed.

Why was it necessary to change? Randomised trials typically include all randomised subjects in the analysis, regardless of the actual receipt of the intervention, in line with the intention-to-treat principle [7,8]. In practice, when applying the intention-to-treat principle, challenges are encountered. First, the intention-to-treat analysis does not fit all scientific questions of interest. Second, if an intention-to-treat analysis is targeted, events after randomisation may preclude observation of the outcomes (e.g. death) or complicate their interpretation (e.g. use of rescue medication). As a partial solution, the intention-to-treat analysis is often complemented by a per protocol analysis, to estimate the treatment effect in patients who adhered to the protocol. Although this estimate may be of interest, selection bias and confounding may be introduced, and the target population cannot be defined outside the conditions of the trial.

Another concern is that the method used for handling missing data is sometimes disconnected from the objective. One of these frequently used methods, a mixed model [8], assumes that patients who are lost to follow-up will have similar outcomes as comparable patients (in terms of the covariates in the model) who did not discontinue the study. This assumption may be unrealistic and as a result may target a different scientific question than the one intended.

The addendum proposes to rethink the intention-to-treat and per protocol analyses. It aims to facilitate the dialogue between disciplines involved in clinical trial planning, design, conduct, data analysis, and interpretation of the results, regarding the treatment effects that the trial seeks to address.

As outlined in the addendum, an estimand defines the target of estimation for a particular trial through specification of the treatment, population, the variable (or outcome), the strategy to handle intercurrent events, and the population-level summary. Intercurrent events are defined as events that occur after treatment initiation and either preclude observation of the variable (e.g. death) or affect its interpretation (e.g. treatment switching).

The addendum suggests five strategies for addressing intercurrent events. In a treatment policy strategy, the value for the variable of interest is used regardless of whether or not the intercurrent event occurs. In a composite strategy, the intercurrent event becomes a component of the outcome. In a hypothetical strategy, a scenario is envisaged in which the intercurrent event would not occur. The fourth and fifth strategies are the principal stratum and while on treatment strategies.

For all key intercurrent events, a strategy should be specified. One analysis combines the different strategies to account for intercurrent events. For example, for one estimand, the patient outcome may be used regardless of the use of rescue medication (treatment policy strategy), while at the same time a patient who drops out due to lack of efficacy will be considered as treatment failure in a composite outcome (composite strategy). An example of implementation of estimands in the PIONEER 4 study is shown in Box 1.

The estimands framework will impact all stages of clinical trials. At the design stage, the protocol should describe key intercurrent events and define the strategy to handle each of them. The prespecified analysis method should align with the trial estimand and make explicit how the intercurrent events and missing data are handled. During data collection, sufficient detail on potential intercurrent events needs to be captured, wherever possible. The estimand also sets requirements on the necessary duration of follow-up. Reporting of trials should be aligned to the estimand framework.

Full implementation of the framework will take time. Although the addendum provides a good starting point, concrete examples of estimands preferred by authorities are still largely unavailable. A plethora of estimands may be proposed by different stakeholders, which may temporarily complicate interpretation and comparison between trials and interventions. A thorough discussion between all stakeholders may harmonise the preferred estimand in specific settings. However, we should protect ourselves from going back to a one-size-fits-all solution. As Einstein once said ‘‘everything should be made as simple as possible, not simpler.’’ We advocate a continued discussion in the literature to continue learning and ensure that implementation of the estimands framework leads to improvement.

What is new?

Key findings

- The addendum "Estimands and sensitivity analyses," was added to ICH E9 guideline "Statistical Principles for Clinical Trials."

What this adds to what was known?

- The addendum proposes to rethink the intention-to-treat and per protocol analyses.
- Using estimands, it aims to facilitate the dialogue between all disciplines involved in clinical trials, regarding the treatment effects that the trial seeks to address.
- An estimand defines the target of estimation for a particular trial through specification of the treatment, population, the variable (or outcome), the strategy to handle intercurrent events, and the population-level summary.

What is the implication and what should change now?

- Adoption of the estimands framework by regulatory authorities and the pharmaceutical industry is expected to lead to global changes in trial design, trial conduct, and interpretation of results.
- We advocate a continued discussion in the literature to continue learning and ensure that implementation of the estimands framework leads to improvement.

Box 1 Example of implementation of estimands

Implementation of an estimand based on a treatment policy strategy for both intercurrent events and another one based on a hypothetical strategy in the PIONEER 4 trial: "Two different questions related to efficacy were addressed through the definition of two estimands: the treatment policy estimand and the trial product estimand. The treatment policy estimand (primary estimand) assessed the treatment effect for all participants randomly assigned to treatment regardless of study drug discontinuation or use of rescue medication. This estimand reflects the effect of initiating treatment with oral semaglutide compared with initiating treatment with subcutaneous liraglutide or placebo, all potentially followed by either discontinuation of study drug or addition of or switch to another glucose-lowering drug, or both. The trial product estimand (secondary estimand) assessed the treatment effect for all participants randomly assigned to treatment under the assumption that all participants remained on study drug for the entire planned duration of the trial and did not use rescue medication. It aims to reflect the effect of oral semaglutide compared with subcutaneous liraglutide or placebo without the confounding effect of rescue medication. The statistical analysis applied with this estimand is similar to how many phase 3a diabetes trials have been assessed in the past. Further details can be found in the full manuscript of the PIONEER 4 study [2].

References

1. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Final version, adopted November 20, 2019. Available at: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf. Accessed June 16, 2020.
2. Pratley R, Amod A, Hoff ST, Kadowaki T, Lingvay I, Nauck M, et al. Oral semaglutide versus subcutaneous liraglutide and placebo in type 2 diabetes (PIONEER 4): a randomized, double-blind, phase 3a trial. Lancet 2019;394:39-50.
3. Pieber TR, Bode B, Mertens A, Cho YM, Christiansen E, Hertz CL, et al. Efficacy and safety of oral semaglutide with flexible dose adjustment versus sitagliptin in type 2 diabetes (PIONEER 7): a multicentre, open-label, randomised, phase 3a trial. Lancet Diabetes Endocrinol 2019;7:528-39.
4. Mosenzon O, Blicher TM, Rosenlund S, Eriksson JW, Heller S, Hels OH, et al. Efficacy and safety of oral semaglutide in patients with type 2 diabetes and moderate renal impairment (PIONEER 5): a placebo-controlled, randomised, phase 3a trial. Lancet Diabetes Endocrinol 2019;7:515-27.
5. Aroda VR, Rosenstock J, Terauchi Y, Altuntas Y, Lalic NM, Morales Villegas EC, et al. Pioneer 1: randomized clinical trial of the efficacy and safety of oral semaglutide monotherapy in comparison with placebo in patients with type 2 diabetes. Diabetes Care 2019;42:1724-32.
6. Mosteller F, Tukey JW. Data analysis, including statistics. In: Jonas LV, editor. The collected works of John W. Tukey: philosophy and principles of data analysis 1965e1986, IV. Boca Raton, FL: CRC Press; 1987:601-720.
7. Hernan MA, Scharfstein D. Cautions as regulators move to end exclusive reliance on intention to treat. Ann Intern Med 2018;168:515-6.
8. Fletcher C, Tsuchiya S, Mehrotra DV. Current practices in choosing estimands and sensitivity analyses in clinical trials: results of the ICHE9 survey. Ther Innov Regul Sci 2017;51:69-76.

CHAPTER 3

A narrative review of estimands in drug development and regulatory evaluation: old wine in new barrels?

Mitroiu M, Oude Rengerink K, Teerenstra S, Pétavy F, Roes KCB. A narrative review of estimands in drug development and regulatory evaluation: old wine in new barrels? Trials. 2020 Dec;21(1):671. https://doi.org/10.1186/s13063-020-04546-1

Abstract

Background: An estimand defines the target of estimation for a clinical trial through specification of the treatment, target population, variable, population-level summary and of the strategies for intercurrent events. A carefully defined estimand aligns the clinical trial design and analysis with the scientific question of interest and adequately accounts for so-called intercurrent events. The ICH E9(R1) addendum suggests five estimand strategies. We evaluated to what extent current practice in drug development and regulatory assessment fits in the estimand framework.

Methods: We systematically evaluated what estimands, especially what strategies for intercurrent events are advised in European Medicines Agency disease guidelines, used in sponsors' trials and additionally requested by the European Medicines Agency in assessment dossiers. We selected four therapeutic areas: nervous system, oncology, cardiovascular diseases and respiratory diseases. For each, we evaluated all disease guidelines with approved drugs, the dossiers of the most recently approved drugs matching the guidelines and corresponding regulatory questions.

Results: Strategies for intercurrent events were present in 18 (53%) of 34 guidelines, in all 34 sponsor documentations and in 15 (44%) of 34 sets of regulatory questions. Treatment policy was advised in 13 (38%) guidelines and was applied in 9 corresponding sponsor documentations. Of these 9, it was the sole strategy in 4 cases and accompanied by another strategy in 5 cases. Hypothetical strategy was not advised in guidelines. However, it was the leading strategy applied in 25 (74%) sponsor documentations. Composite strategy was advised in 3 (9%) guidelines and applied accompanied by another strategy in 2 corresponding sponsor documentations. While on treatment strategy was not advised in guidelines, but was applied in 2 sponsor documentations. Principal stratum strategy was advised in 2 guidelines but not applied in corresponding sponsor documentations. Of the regulatory questions, treatment policy was present in 2 cases (6%), hypothetical in 6 cases (18%), composite in 6 cases (18%) and while on treatment in 1 case (3%).

Conclusions: Estimand attributes are present in guidelines, sponsor documentations and regulatory questions, but not described as estimands. Treatment policy was most often advised in guidelines, but hypothetical was the leading strategy applied in sponsor documentations. Thus, results indicate not a full concordance between the regulatory target of estimation and what is actually estimated. The lack of concordance was mostly due to limitations in collection of intercurrent events data to enable a treatment policy strategy. There is, therefore, a need to better define estimands at the design stage and throughout the applications dossiers and assessment reports.

Background

In randomised controlled clinical trials, the aim is to estimate the effect of an intervention compared to a control treatment, unconfounded by assignment to intervention or control. Through randomisation, it is intended that any difference in clinical outcome can be attributed to the intervention and can be causally interpreted [1]. In practice, post-randomisation events, such as treatment discontinuation, use of concomitant medication or a switch in treatment arm, may be related to the treatment. These post-randomisation events cause missing outcome values or more complex; they introduce bias in outcomes obtained. They do not preserve randomisation and subsequently do not allow the randomisation-based inference, hence impacting the estimation of the treatment effect and/or its interpretation. In this context, many methods were proposed to deal with missing data, such as mixed models or imputation methods, or using a composite endpoint treating missing values as non-responders in order to minimise bias [2–4]. However, little attention was given to what impact these missing data handling methods actually have on the treatment effect to be estimated at target population level in realistic conditions. There was a fundamental lack of common understanding between involved stakeholders of what these methods aim to estimate in relevant target patient population terms, as well as a lack of harmonisation in applying and interpreting these methods [5].

The ICH E9 [6] describes the intention-to-treat (ITT) principle and the analysis set (FAS).

> *"The intention-to-treat principle implies that the primary analysis should include all randomised subjects. Compliance with this principle would necessitate complete follow-up of all randomised subjects for study outcomes. In practice this ideal may be difficult to achieve, for reasons to be described. In this document the term 'full analysis set' is used to describe the analysis set which is as complete as possible and as close as possible to the intention-to-treat ideal of including all randomised subjects."*

Hence, it points out to the reader that in practice, it may not be possible to have all outcomes observed for all randomised patients in order to comply with the intention-to-treat principle. Furthermore, the full analysis set is used to describe the population almost the same as all randomised patients and certain criteria are mentioned (with respect to treatment intake and missing data) that could lead to patients being excluded from the FAS, e.g. "the failure to take at least one dose of trial medication and the lack of any data post randomisation". However, it does not mention the scenario when some of the randomised patients have only partially observed outcome data, such as in a longitudinal trial with repeated measurements at protocolled visits, that have all visits but the

last one at end of trial, irrespective of other post-randomisation events that did not lead to missing data. Therefore, the trialists are in a difficult and challenging position where something has to be done for the patients with partially or fully missing outcomes (e.g. after they discontinue study, regardless of their reason) in order to comply with the intention-to-treat principle and to reach a full analysis set. It is yet unclear what was done or what can be done in order to include these patients in the (m)ITT analysis.

The term "estimand" is not new in statistics [7]. More recently, it was used as a solution for the "missing data problem" [8,9]. In 2017, it was incorporated into ICH E9(R1) draft addendum on estimands and sensitivity analysis in clinical trials [10,11], primarily to precisely define the treatment effect in a randomised trial. This addendum supplements the ICH E9 guideline "Statistical Principles for Clinical Trials" [6] from the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) [12]. The addendum recommends that the estimand should be precisely defined upfront, which addresses more than the "missing data problem".

The draft E9(R1) addendum defines four attributes to describe the estimand: variable (or outcome), population, population-level summary and strategies to account for intercurrent events.

The variable (or outcome) to be obtained or measured for each individual patient that is required to address the scientific question. If we use an example from pain medication, the variable could be a visual analogue score (VAS) obtained at pre-specified visit times in a trial for acute pain treatment, e.g. VAS to be measured or obtained at baseline, at week 4, week 8 and at week 12 (end of trial).

The population, referring to the patients targeted by the scientific question. In the trial for pain treatment, the population could be "adults suffering from acute pain".

The population-level summary for the variable which provides a basis for a comparison between treatment conditions. For example, it could be the difference in VAS means between the experimental and control arm at a pre-planned timepoint, e.g. at 12 weeks.

The specification of how to account for intercurrent events to reflect the scientific question of interest (through strategies for intercurrent events); intercurrent events are defined in E9(R1) addendum as "events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest." The E9(R1) addendum suggests five strategies to address intercurrent events: (1) "treatment policy", (2) "hypothetical", (3) "composite",

(4) "while on treatment" and (5) "principal stratum". For instance, in the trial for pain treatment, self-administration of additional medication for pain might be prohibited by the protocol, but some patients do take it. With a treatment policy strategy, the intercurrent event "need for additional medication for pain" is actively ignored, and the VAS is used as it is for those patients that take additional medication. The treatment policy strategy would technically correspond to the intention-to-treat principle. With a hypothetical strategy, a scenario is envisaged where the intercurrent event "need for additional medication for pain" would not occur. With this strategy for instance, the VAS values following intercurrent event are set to missing if such is in accordance with the hypothetical scenario considered. With a composite strategy, the intercurrent event is explicitly taken into account and made part of the outcome, for instance, by assigning a worst value of VAS, or by considering the patient a non-responder if a binary outcome is used. With a while on treatment strategy, for this intercurrent event takes the form "while no need for additional medication for pain" and VAS values following intercurrent event are not of interest. With a principal stratum strategy, based on baseline covariates, the stratum of patients that would not experience the intercurrent event is tried to be identified. Analysis is then conducted on this stratum. The addendum informs that principal stratum should be distinguished from any type of analysis in a subgroup of patients, such as per protocol or complete case analysis. The E9(R1) addendum also describes scenarios with two different intercurrent events handled by the same strategy or each of the two intercurrent events handled by a different strategy.

The final version of the ICH E9(R1) addendum was published in December 2019 and uses five attributes [13]. One of the five attributes from the final addendum, the "treatment", was added compared to the four attributes of the draft addendum. The strategies for intercurrent events and their definitions are not different between the draft and the final versions of the addendum. The other three attributes were slightly restructured. In the remainder of this article, we followed the structure and the four attributes from the draft version of ICH E9(R1) addendum.

It was expected that the estimand was not defined explicitly in the terms of these attributes in protocols and reports before publication of the draft addendum. However, clinical trials still had a primary objective with a primary outcome variable, a target of estimation at population level and a pre-specified statistical analysis. This entails that to some extent and at least implicitly, the key elements of an estimand are expected to be present in clinical trials before the E9(R1) addendum concepts became public.

A survey published in 2017 found that an intention-to-treat estimand was most often aimed at and that the most often used methods for missing data handling were

mixed-models repeated measures (MMRM) or last observation carried forward (LOCF) imputation [14]. In the precise language of E9(R1), there is likely a mismatch between the aim of intention-to-treat ("treatment policy") and these often used methods of dealing with missing data. Hence, impact of implementation on design and analysis of trials can certainly be expected, but it is currently unclear how large the impact of the proposed estimand framework may be. It is important to identify to what extent the framework leads to different effect estimates compared to current practice in drug development and regulatory assessment.

We therefore aimed to answer the following research questions:
1. What types of estimands, especially what strategies to account for intercurrent events, are advised in European Medicines Agency (EMA) disease guidelines?

2. What types of estimands, especially what strategies to account for intercurrent events, are used by sponsors in their confirmatory clinical trials supporting the application for marketing authorisation?

3. What additional types of estimands, especially what strategies to account for intercurrent events, are requested by the regulatory agencies in reply to the sponsor documentations in assessment dossiers?

Methods

We systematically evaluated what estimands were targeted in regulatory disease guidelines, in trials from recently approved applications and in regulatory questions. We scrutinised what strategies to account for intercurrent events were advised, used and further requested in drug development and evaluation. We performed this review on EMA [15] disease guidelines [16] and on corresponding approved medicinal products dossiers [17].

Selection of disease guidelines and medicinal products for evaluation

First, we selected all EMA disease guidelines (described hereafter as the "guidelines") within four therapeutic areas: nervous system, oncology, cardiovascular diseases and respiratory diseases, to identify the diseases for which regulatory guidance is available for clinical efficacy investigation. We considered these four main therapeutic areas to have the broadest coverage of most estimands practices.

In November 2017, for all identified diseases within these four therapeutic areas for which regulatory guidance is available, we selected the most recently approved innovative product in the centralised procedure (Figs. 1 and 2), defined by the date of positive opinion from the Committee for Medicinal Products for Human Use (CHMP) [18].

The most recently approved products were assumed to best reflect current practice. We limited our selection to one product within each disease as a snapshot of how the estimands principles were employed in practice. We excluded guidelines for which there was no approved product available up to November 2017. We used the version of the guideline that was effective at the time of approval for each particular medicinal product, with a few exceptions. In case a new guideline became effective closely after product approval, it was assumed that draft information was available through the public consultation phase, therefore impacting already clinical trial design in practice.

For each selected product, we used the sponsor's protocols, statistical analysis plans and clinical study reports (altogether described hereafter as the "sponsor documentation") for the confirmatory clinical trials supporting the application for authorisation. We extracted the questions raised by the EMA during the assessment procedure verbatim from the "day 120 list of questions" of the centralised procedure (described hereafter as the "regulatory questions"). This list of questions is expected to capture the

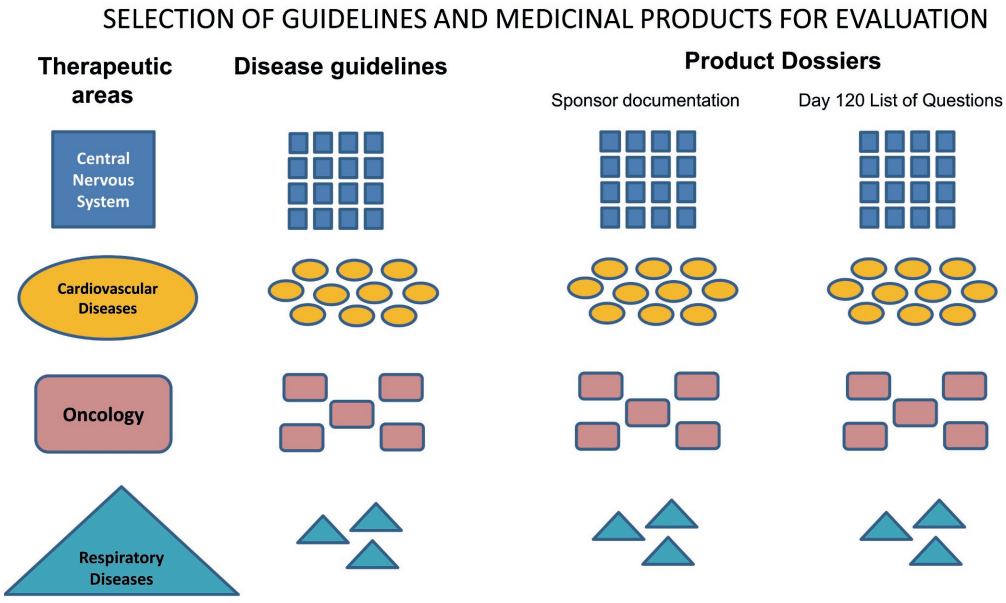


Figure 1. Illustration of the process employed for data extraction and interpretation

most extensive and least selective list of efficacy-related questions raised by the EMA. Regulatory questions contain the "major objections" and the "other concerns", which can be supplementary questions addressed by regulators to the sponsor, based on the evidence provided in the application dossier which includes the sponsor's documentation [19].

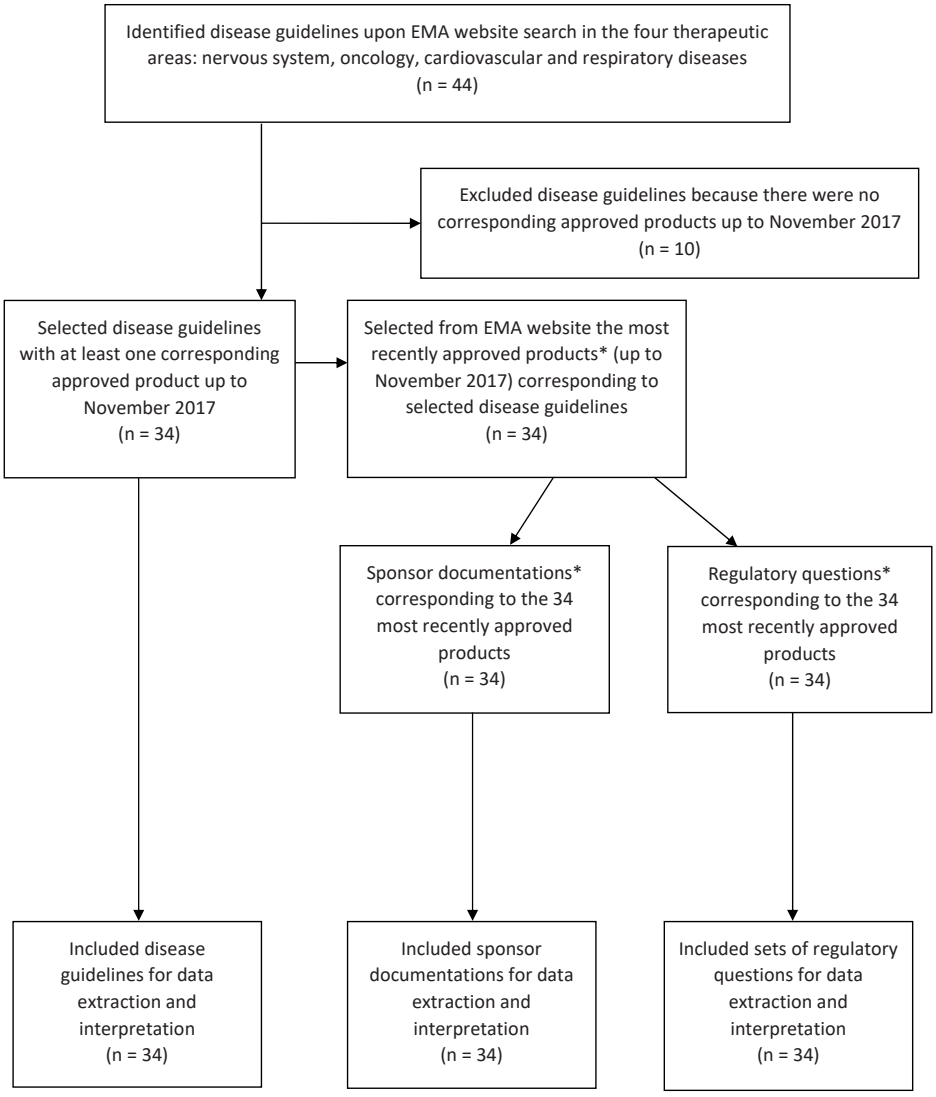


Figure 2. Flow diagram detailing the identification, selection and inclusion of disease guidelines and approved products for data extraction and interpretation

**one approved product= one sponsor documentation and one corresponding set of regulatory questions*

Data extraction

We used three sources of information for data extraction: the guidelines, the sponsor documentation and regulatory questions. We extracted from each data source all relevant phrases and paragraphs pertaining to the target of estimation (estimand and its attributes as per ICH E9(R1) draft addendum) corresponding to the primary analyses and corresponding supplementary/sensitivity analyses:

- Variable/outcome,
- Population,
- Factors that are likely to influence the treatment effect (e.g. rescue medication) (potential intercurrent events) and information regarding the missing data handling,
- Comparison (statistical contrast) upon which treatment effect is interpreted, called population-level summary in the ICH E9(R1) addendum.

For each selected disease guideline, we used the most recent version from the EMA website. For each selected medicinal product, we used the dossiers that we retrieved from the document management system of CBG-MEB and EMA database. We created a data extraction form, collected and stored the information in Microsoft Office Access Database (Appendix 2).

We pilot-tested the first version of the data extraction form on two products for guideline, sponsor documentation and regulatory questions. Following this pilot test, we refined the form.

Data interpretation and translation from efficacy-related information to estimand attributes

During the research, it became obvious it is necessary to add a data interpretation step in order to translate the raw, unstructured information derived from the documentation into estimand constructs. This was the case especially for the intercurrent events and strategies to account for them. The information pertaining to variable, population-level summary and target population was more straightforward to map from the basic information.

For each guideline, corresponding sponsor documentation and regulatory questions respectively, we interpreted the estimand attributes in order to reconstruct the corresponding implied estimand.

The E9(R1) draft addendum was not yet published when the guidelines were published, when the trials were conducted or when the regulatory questions were raised; therefore, the attributes were not expected to be phrased and framed in a dedicated section and

not in the shape specified in E9(R1) draft addendum. The information had to be translated from text referring to the efficacy analysis into an estimand attribute (e.g. "concomitant medication use" as intercurrent event). For population-level summary, we used the comparison (statistical contrast such as difference in means, odds ratio) on which the treatment effect quantification and/or magnitude is assessed and concluded on. This was typically derived from the proposed primary analysis. For population, we used the population description in analyses or in analyses sets. The variable could be extracted as it was described. We categorised the attributes as "present" or "not present". The attributes are qualified as "present" if they fulfil the above definitions as per draft E9(R1), regardless of where in the documents the information was found. In the example below, the attributes were considered "present".

Attribute	Variable	Population	Strategy to account for the intercurrent event	Population-level summary
Phrase verbatim extracted	"the primary endpoint for the primary analysis is LDL-c"	Inclusion/exclusion criteria + "The FAS comprised all patients that were randomised and had an evaluable outcome value at 12 months"	"patients with missing values that switch regimens or discontinue assigned treatment are counted as failures"	"Difference in mean change from baseline of LDL-c at 6 weeks"

Conversely, if only non-specific statements, such as "The effect of missing values will need to be taken into account in the efficacy analysis and the method to address this problem needs to be pre-specified", without being incorporated in the analysis or without clear and explicit instruction, then the intercurrent event attribute is considered "not present". All other attributes will be qualified as "not present" if they are not specified and cannot be determined given their definitions in E9(R1) draft addendum.

Within intercurrent events, we created the category of intercurrent events "not accounted for" (NAF). This category represents the intercurrent events that were identified and possibly collected, but not included or referred to in the primary efficacy analyses, for example, "concomitant administration of systemic corticosteroids", "change in background medication" or "salty food intake".

We determined whether the estimand could be reconstructed from the information provided. If all attributes are scored "present", then the estimand is classified as "can be determined". If any of the attributes are deemed "not present", then the implied estimand will be classified as "cannot be determined". We interpreted the strategies using the five types of strategies proposed and defined in E9(R1) addendum: treatment policy, hypothetical, composite, while on treatment and principal stratum. Where the

strategies type did not fit in one of the E9(R1) definitions, we described the strategy in detail and classified them as "other". Per protocol analysis was not defined in the addendum, but depending on how it is defined then it could have been correspondent to a strategy, e.g. while on treatment.

To understand the strategies for intercurrent events and reconstruct the estimand, we also extracted information pertaining to the statistical analysis and imputation methods.

Quality review

The concept of intercurrent events is the novel aspect introduced with the estimand framework. The quality review therefore focused especially on the strategies to account for the intercurrent events, as these needed most interpretation. The other estimand attributes were deemed unambiguous to determine by the primary data extractor and interpreter (primary reviewer) as well as by the secondary reviewers. As we found that guideline texts often led to difference in opinions regarding presence of intercurrent events and strategies to account for them, we proceeded with full double review of all guidelines. One person (MM—primary reviewer) extracted the data, translated it to estimand attributes and reconstructed the implied estimand. For quality control of all guidelines estimand constructs, two secondary reviewers (ST and KOR) each read the entire guidelines content and reviewed the estimands constructs next to the primary reviewer. Differences were solved in consensus between the primary and secondary reviewers.

If consensus could not be reached between the primary and secondary reviewer, a third reviewer would be consulted (KR/FP) and discussed until consensus was reached. Each secondary reviewer performed the quality review for half of the selected set of guidelines.

The sponsor documentations and regulatory questions were less ambiguous to interpret, and we considered the efficacy analyses in general detailed enough in order to adjust the depth of the quality review. Each secondary reviewer performed the quality review independently for five different products dossiers (sponsor documentations and the corresponding regulatory questions). If > 25% of the intercurrent events and implied strategies would not match, then a full review would be triggered for all sponsor documentations and corresponding regulatory questions. If < 25% discrepancies but with systematic or recurrent errors (e.g. consistent mismatch in a particular strategy or combination of strategies), the primary reviewer would re-review all dossiers for that particular error.

Analysis and summary of results

We summarised the estimands and attributes overall and per therapeutic area. The experimental unit for analysis was considered a guideline, the set of sponsor documentation pertaining to one product or the set of regulatory questions related to the sponsor documentation corresponding to that product, respectively. We created cross-tabulations for attributes ("present" or "not present"), intercurrent events and type of strategy for intercurrent events according to the types proposed in E9(R1) draft addendum. We summarised what strategies for intercurrent events and in which combination they were used with the other attributes to define the estimands. If no strategy or no estimand was specified or described, these were summarised as "strategy not present" or "no estimand present". Given the nature of the review and summaries, no statistical testing was performed.

Results

We included 34 guidelines for which products were approved, 34 sponsor documentations for the approved products corresponding to the guidelines and 34 sets of regulatory questions corresponding to the approved products we had selected (Appendix 3). Those were selected from the therapeutic areas nervous system (n = 16), oncology (n = 5), cardiovascular diseases (n = 10) and respiratory diseases (n = 3). Guidance documents effective dates ranged from 1992 to 2017, products approval ranged from 1996 to 2017 and regulatory questions dates ranged from 1995 to 2016. Two secondary reviewers agreed without or with limited changes with the data extracted and their interpretation. For the 10 sponsor documentations and regulatory questions that were reviewed in pairs, reviewers agreed more than 75% of extractions (90% with ST, 80% with KOR); hence, full peer review of all sponsor documentations and regulatory questions was not triggered.

Description of the four estimand attributes

All four estimand attributes were specified in 12% of the guidelines, in all sponsor documentations and in 3% of the regulatory questions (Table 1). We found the information pertaining to attributes scattered in different sections throughout the statistical analysis plans, protocols and clinical study reports. The information pertaining to attributes was easy to find in guidelines, but more difficult to find in sponsor documentations. The attributes were not described explicitly and often embedded in primary efficacy and statistical methods, missing data handling, data collection or results sections. If described, the attributes were found relatively easy in the regulatory questions in the section for clinical efficacy (Appendix 1). However, not all attributes are described explicitly for all analyses requested in the regulatory questions.

Table 1. Frequency of attributes presence/description in guidelines, sponsor documentations and regulatory questions

	Frequency of attribute presence		
Source document Attribute	**Guidelines (N=34) (%)**	**Sponsor documentations (N= 34) (%)**	**Regulatory questions (N=34*) (%)**
Variable	100	100	68
Population	24	100	3
Population-level summary	38	100	3
Intercurrent events	79	100	68
Strategies intercurrent events	53	100	44

**23 out of 34 regulatory questions documents had estimand-related questions*

The variable (or outcome)

The variable was present in all 34 guidelines, in all 34 sponsor documentations and in 23 (68%) of 34 sets of regulatory questions.

The population

The population was described in 8 (24%) guidelines, in all 34 sponsor documentations and in 1 (3%) set of regulatory questions. An intention-to-treat analysis (ITT corresponding to a treatment policy strategy) is advised in most guidelines. We found in sponsor documentations multiple ways in which the analysis population deviated from the definition of the intention-to-treat principle [6]. Even if described using the same term "modified intention-to-treat", the modifications varied between products and studies (Table 2). And, although the term "Intention-to-treat" was used, not all randomised patients were included in the analysis as the ITT principle dictates.

The population-level summary for the variable

The population-level summary was present in 13 (38%) guidelines, in all 34 sponsor documentations and in 1 (3%) set of regulatory questions.

The strategies to account for intercurrent events

Intercurrent events were described in 27 (79%) guidelines, in all 34 sponsor documentations and in 23 (68%) sets of regulatory questions. Strategies to account for intercurrent events were present in just over half of the guidelines (n = 18, 53%), in all sponsor documentations (n = 34, 100%) and in almost half of the regulatory questions (n = 15, 44%) (Table 1).

Treatment policy strategy was advised in 13 (38%) guidelines. It was applied in 9 corresponding sponsor documentations. Of the 9, it was the only strategy applied in 4, and in 5 it was applied accompanied by another strategy or analysis for different intercurrent

events. In the four remaining sponsor documentations, a different strategy (or a mix of other strategies) than treatment policy was applied. Treatment policy strategy was applied in total in 13 sponsor documentations; hence, in 4 of these 13 sponsor documentations, it was applied in the absence of being advised in the guideline.

Hypothetical strategy was not advised in any of the 34 guidelines. However, it was the sponsors' preferred strategy, applied in 25 sponsor documentations, especially to account for missing outcome values (caused by known or unknown intercurrent events). Hypothetical strategy was generally used in the same estimand simultaneously with another strategy, usually with the treatment policy strategy applied for a different intercurrent event. The typical hypothetical strategies were identified in relation to LOCF, MMRM and censoring in time-to-event analysis. These were related to missing data and were used as a measure to explicitly/implicitly impute or handle missing outcomes that were planned to be collected but were not.

Composite strategy was advised in 3 (9%) guidelines. It was applied in 2 corresponding sponsor documentations. Of the 2, it was not applied as single strategy in any of the corresponding sponsor documentations; it was applied accompanied by another strategy or analysis for different intercurrent events. In the one remaining corresponding sponsor documentation, a different strategy (or a mix of other strategies) than composite was applied. Composite strategy was applied in total in 6 sponsor documentations; hence in 4 of these 6 sponsor documentations, it was applied in the absence of being advised in the guideline.

While on treatment strategy was not advised in any guideline but was applied in 2 (6%) sponsor documentations. Clinical outcome (events) was measured over the non-missing days or number of events were adjusted for the treatment period (a negative binomial model with offset for treatment exposure period), both in CNS therapeutic area. The population-level summary was a contrast for rates of events.

Principal stratum strategy was advised in 2 (6%) guidelines but was not applied in any sponsor documentations.

Table 2. Variations of intention-to-treat (ITT)

Population description	Name	Formulation(s)
ITT, mITT, FAS	Intention-to-treat, Modified intention-to-treat, Full analysis set	1. All randomised patients with at least one follow-up measurement available
		2. All randomised patients that took any/at least one dose of trial medication
		3. All randomised patients with the baseline measurement available and at least one post-baseline measurement available
		4. All randomised patients with baseline measurement available, at least one post-baseline measurement and took any/at least one dose of trial medication
		5. All randomised patients with at least one post-baseline measurement and took any/at least one dose of trial medication

The estimands advised in guidelines and requested in regulatory questions contained a single strategy intending to cover one or multiple different intercurrent events at the same time, such as a treatment policy strategy applied for all intercurrent events. The estimands in sponsor documentations contained multiple strategies to concomitantly handle multiple different intercurrent events at the same time, such as a treatment policy strategy applied for one intercurrent event and hypothetical strategy for another.

In the 16 sponsor documentations corresponding to guidelines where a strategy was not advised, the hypothetical strategy was most often used for some intercurrent events and in conjunction with another strategy for other intercurrent events.

Of the regulatory questions, treatment policy strategy was present in 2 (6%), hypothetical strategy in 6 (18%), composite strategy in 6 (18%), while on treatment strategy in 1 (3%) and principal stratum strategy in none (0%).

Apart from the five strategies suggested and defined in the draft E9(R1), we found other types of analyses that do not fall within any of the five strategies definitions, we summarised these as “other” (Fig. 3). We found them in guidelines, sponsor documentations and regulatory questions. Over half of the “other” category was a per protocol analysis, using various definitions of protocol violations or deviations. Furthermore, we also encountered complete cases or available case analyses. None of these could be usefully categorised as estimand strategy.

Strategies by therapeutic area

Treatment policy was the most often advised strategy in guidelines for each therapeutic area. Composite and principal stratum strategies were present in central nervous system and cardiovascular disease guidelines but were not present in guidelines for oncology and respiratory diseases (Fig. 4a).

The estimands suggested in guidelines contained a single strategy covering multiple different intercurrent events at the same time, such as a treatment policy strategy applied for all intercurrent events.

In sponsor documentations, hypothetical strategy was the leading strategy in each therapeutic area, followed by treatment policy strategy. While on treatment strategy was seen only in cardiovascular diseases. All therapeutic areas, except respiratory diseases, used the composite strategy (Fig. 4b). In sponsor documentations, they usually aimed at a treatment policy estimand. This, however, is often not strictly achievable as per ICH E9 mainly due to limitations in the data. Reported estimands contained multiple different strategies for different intercurrent events or other analyses. For example, they applied within the same analysis treatment policy for some intercurrent event(s) and a hypothetical strategy for other intercurrent events that led to missing data.

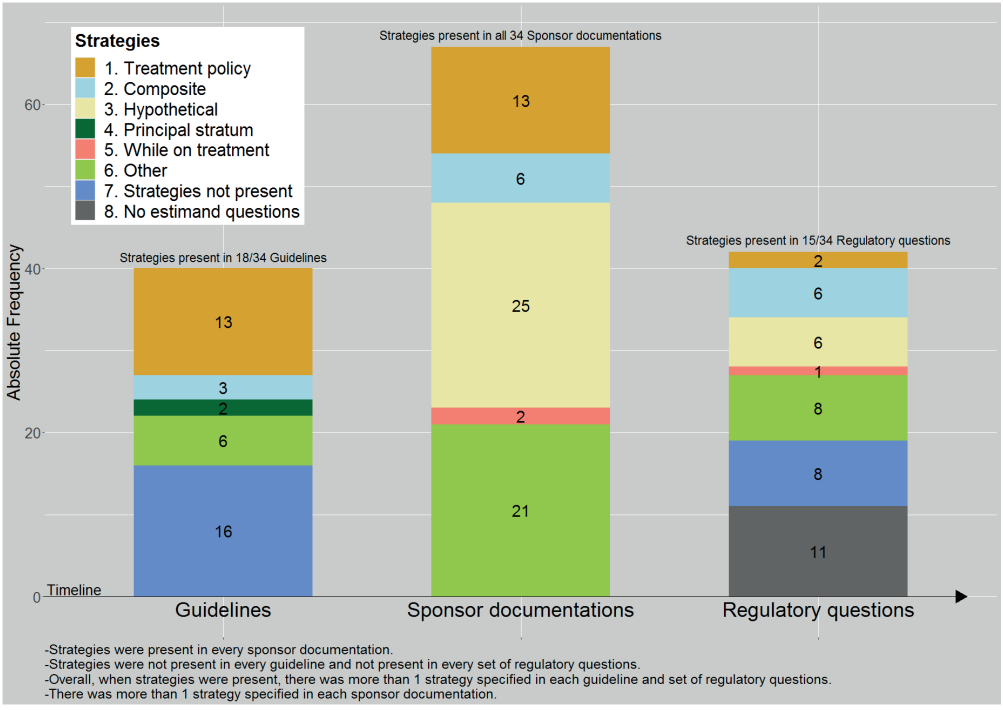
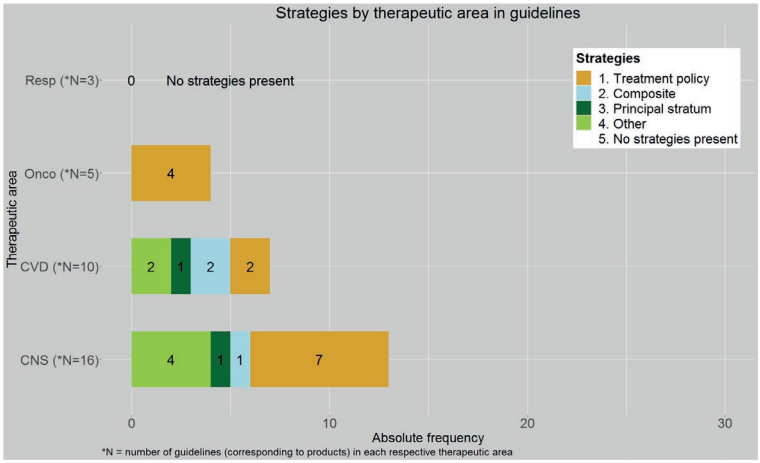
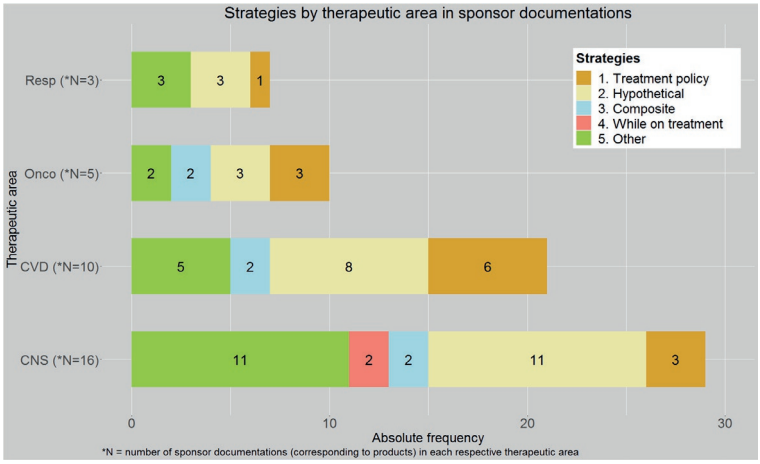


Figure 3. Stacked barplot with strategies in guidelines, sponsor documentations and regulatory questions

a



b



c

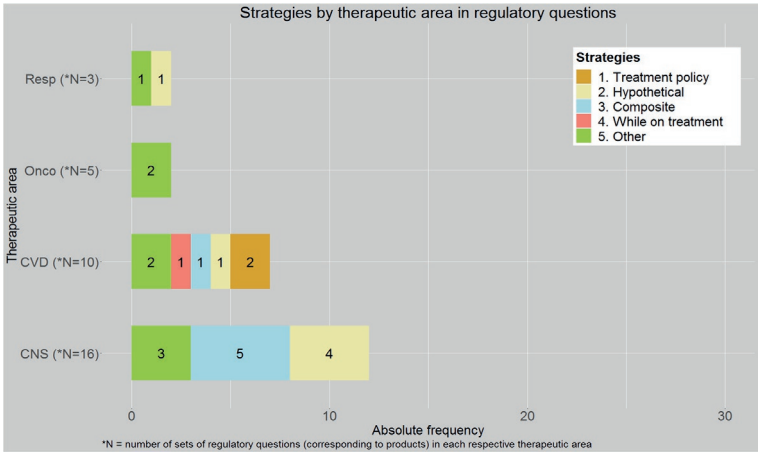


Figure 4. Strategies by therapeutic area in a guidelines, b sponsor documentations and c regulatory questions. CNS, central nervous system; CVD, cardiovascular diseases; Onco, oncology; Resp, respiratory diseases

Regulatory questions typically contained one estimand (strategy) or analysis per question. The strategies or analyses requested were usually different from a treatment policy strategy, with no clear different pattern between therapeutic areas (Fig. 4c).

Discussion

With this review, we provide an overview of what the implied estimand practices were in drug development and regulatory evaluation before the publication of the draft version of ICH E9(R1) estimand framework (Fig. 3). Sponsor documentations contained more detailed descriptions of the estimand attributes than guidelines and regulatory questions. A treatment policy strategy was most often advised in guidelines and targeted in sponsor documentations. However, a treatment policy strategy could often not be fully achieved due to incomplete follow-up, resulting in a hypothetical estimand being the most frequent approach by sponsors. Apart from the five strategies defined in the addendum, we also identified other analyses types.

The variable was the estimand attribute most present and clearly defined in guidelines, sponsor documentations and regulatory questions. This was not surprising, as the clinical outcome to be obtained or measured in patients is a pivotal item to decide on, when designing a trial. It already is thoroughly discussed between involved parties. Hence, it is usually described in detail and concordant in all types of documents.

In all sponsor documentations, we found data collected and reported, for example, for drop-out due to adverse events or concomitant medication and these data are used for instance, in safety analyses. However, the included information on intercurrent events (such as an adverse event leading to study withdrawal) was not used nor referred to in the primary efficacy analysis.

We found that strategies advised in guidelines, applied in sponsor documentations and asked for in regulatory questions, were different. There could be several reasons behind this finding; it could be due to the fact that sponsors may have followed advice from disease guidelines under the remit of other regulatory authorities than EMA, such as the FDA guidelines.

Furthermore, guidelines may have advised a single general strategy that cannot be applied for all intercurrent events, such as treatment policy. Sponsors applied the advised strategy for part of the intercurrent events where the advised strategy could be applied, but also had to apply another strategy or analysis for other intercurrent events where the advised strategy could not be applied.

This could be due to lack of precision in articulating the targeted treatment effect in guidelines and also due to feasibility of applying only one strategy given the data and not because of sponsors' lack of intent to apply advised strategies. It may also be the case that sponsors identified different or more intercurrent events than the ones suggested by guidelines and, hence, based on their treatment effect of interest, decided to apply other strategies than the strategies advised in guidelines. EMA can provide sponsors with scientific advice regarding their trials; however, not all sponsors apply for scientific advice. Moreover, having been provided with scientific advice, as it is not legally binding, it does not make it mandatory for the sponsors to follow it [20].

In spite of sponsor documentations providing more detailed specifications of analyses and estimand-related information, it could often not be inferred from the documentation why certain choices were made for specific intercurrent events. Similarly, the clinical questions raised during the assessment were not phrased precise enough to translate into the intended estimand. Therefore, it was not possible in general to determine how well aligned the estimand was with the clinical question of interest.

Although the strategies in guidelines seem to differ between therapeutic areas, the types of strategies found in sponsor documentations seemed to be similar between therapeutic areas.

The difference between strategies advised in guidelines and implemented by the sponsors could be attributed to several reasons. First, strategies advised in guidelines may not be fully achievable in practice. Secondly, sponsors may have followed disease guidelines and feedback from other regulatory regions, such as from FDA, which might have advised different strategies.

Over half of the "other" category was a per protocol analysis, using various definitions of protocol violations or deviations. We also encountered complete cases or available case analyses. However, none of these yields a proper estimate for a meaningful estimand as the results cannot be generalisable to a broader target population. It is debatable if such an analysis is actually targeting estimates in a principal stratum. It would be at very best an improper analysis for it. In this respect, the draft addendum informs that "treatment effects in principal strata should be clearly distinguished from any type of subgroup or per protocol analyses where membership is based on the trial data". To enable analyses in strata, causal inference approaches are likely necessary.

A strength of this research is that this review is the first of this nature. It acts as a snapshot of actual practices with estimands and strategies present in documents pertaining to

medicines development and regulatory evaluation. It was done with access to extensive and full documentation of the actual protocols, results and regulatory interaction. It could be used as baseline in future publications following implementation of ICH E9(R1) addendum. We used publicly available guidelines and had access to regulatory agencies databases. Concomitantly with the publication of the E9(R1) addendum and from its perspective, this review provides valuable insights into current estimand practices. Thus, it can facilitate the implementation of the estimand framework in drug development and regulatory evaluation of medicinal products.

Our review also has some limitations. First, there could be other guidelines in other therapeutic areas or within those therapeutic areas for which products were not approved that contain other strategies for intercurrent events. The results observed might not be fully generalisable to those areas. However, this review represents an overview of the estimand practices in the four biggest therapeutic areas. Second, we had to interpret the intercurrent events and strategies to reverse-engineer the estimand. This process was difficult mainly because the clinical question was not detailed and clear enough to understand the targeted treatment effect and the attributes' information was not structured consistently throughout all documentation. The possible bias due to the interpretation was partly addressed with an independent quality review from two different secondary reviewers which resulted in high percentages of concordance. It is therefore likely that it did not have much impact on the results. Furthermore, we extracted the estimand attributes as defined in the draft addendum and not as in the final addendum. Therefore, we did not collect data for "treatment" attribute. As our attention was focused mainly on the strategies used for intercurrent events, it did not impact our results and conclusions. We were not able to identify clear and unambiguous clinical questions that are addressed in the trials. Most of the times, the clinical trial objective is phrased as "to study the effect of experimental treatment X over control in patients suffering from Y". We consider this to be insufficiently described, and according to the addendum, this is key to enunciate in detail and adequately. This impacts what treatment or treatment strategy is investigated and directly affects the estimand, regulatory evaluation and approval, and ultimately the label of the medicinal product.

The estimand framework is expected to impact all phases of drug development and regulatory evaluation. Defining the estimand aims to provide clarity and better define the treatment effect in perspective to the question of interest. It will consequently facilitate interaction between regulators, patients, clinicians, investigators, HTA bodies, statisticians and other trialists. Therefore, changes are needed for successful roll-out and alignment with the estimand principles. The estimand framework may not solve the causes of trials issues, such as incomplete patient retention or poor treatment

compliance, but it would add clarity on how these can be handled in a transparent and principled manner. Apart from dissemination of the estimand framework in all branches of the medical community by means of training materials, workshops or research articles, another important step might be to update templates of trial protocols. These templates are used at initial stages of development by every involved stakeholder, irrespective of trial type or phase. One such commonly used document where the estimand framework can be introduced is the ICH M11 guideline [21]. This would formalise the need for estimand discussions early in a trial and by all stakeholders. Starting drug development using the estimand framework would ensure that subsequent stages (e.g. study or assessment reports, prescription information for patients etc.) follow the same structure and principles.

In practice, it would be very unlikely that one type of estimand or one type of strategy for intercurrent events would be satisfactory for all stakeholders [22]. For example, a regulatory body might be interested in a treatment policy strategy for an intercurrent event, while a patient would be interested in a principal stratum strategy for the same intercurrent event. It is still to be revealed by further research, how and under which conditions a clinical trial can answer different clinical questions of interest for different stakeholders, with different estimands and/or different strategies for intercurrent events. We suggest to be descriptive and explicit regarding what strategy or strategies are advised, applied or requested, for what intercurrent events. Additional to detailed descriptions, we could use for instance "single-strategy estimand" to define an estimand with one strategy handling one or more intercurrent events at the same time and "multiple strategies estimand" to define an estimand with two/ more different strategies handling two/more different intercurrent events.

Furthermore, we hope the estimand framework is implemented as envisaged in the addendum, to improve the quality with which clinical research questions are addressed by clinical trials. This includes reaching agreement between stakeholders on the estimand(s) of interest, in a transparent, principled and efficient manner.

In ICH E9, the ITT principle is defined [6]. In actual practice, many different deviations from the principle were encountered under the term "modified ITT". However, any modification to ITT definition based on observed trial data (e.g. patients having to take at least one dose of assigned treatment) may not clearly define an actual targeted population anymore and make results difficult to interpret.

Aiming for an estimand does not guarantee estimability of that estimand from the trial data, and we think this is a common pitfall for interpretation of trials. For instance,

treatment policy strategy can be pre-planned to be applied for all intercurrent events but cannot be achieved in full because of missing outcome values for other reasons. ICH E9 informed the reader that the ITT principle may be difficult to achieve as it needs complete follow-up of all randomised subjects for study outcomes [6]. For intercurrent events causing missing data or for missing outcome values, another strategy (e.g. hypothetical or composite) might be applied. Thus, this may lead to an actual estimand that is different from the one aimed at and to answering a question that deviates from the intended clinical question of interest. This situation is encountered in trials and should be acknowledged.

The estimand framework can help in the design of a trial, to pro-actively strike a balance between the estimand aimed for in principle and an estimand that is actually possible to estimate. It will also help revealing the gap between targeted and realised estimands and facilitate discussion among all stakeholders resulting also in a better understanding of drug effect and better comparison across trials or in meta-analysis of clinical trials.

So, are estimands old wine in new barrels? Estimands are both new and old, and missing data as well as intercurrent events in clinical trials are a long existing issue in medical research. Conceptually, it appears the estimands are as "old" as medical research and clinical trials, because it always had estimand elements (e.g. outcome measured) and even if empirically estimated, there was a target of estimation. Estimands in the shape principled by ICH E9(R1) are an innovative solution to deal with fundamental elements of clinical trials, starting from the research question and dealing with intercurrent events, missing data and treatment effect definitions. The estimand framework provides a new framework to align key elements of design, conduct and analysis of clinical trials to adequately answer the clinical question at hand [23].

Conclusions

Estimand attributes are present in guidelines, sponsor documentations and regulatory questions, but not described as estimands. Treatment policy was most often advised in guidelines, but hypothetical was the leading strategy applied in sponsor documentations. Thus, results indicate not a full concordance between the regulatory target of estimation and what is actually estimated. The lack of concordance was mostly due to limitations in collection of intercurrent events data to enable a treatment policy strategy. There is, therefore, a need to better define estimands at the design stage and throughout the applications dossiers and assessment reports.

Acknowledgements

We are deeply thankful to André Elferink, Veronica Quinto, Florian Lasch, Simona Vila, Beatriz Rabazo, Victoria Palmi Reig, Catherine Drai and Ralf Herold for their help and useful comments.

Availability of data and materials

The guidelines used in this research are available from European Medicines Agency website: www.ema.europa.eu

The dossiers data are not publicly available.

References

1. Lachin JM. Statistical properties of randomization in clinical trials. Controlled Clinical Trials. 1988;9:289–311.
2. Liu GF, Pang L. On analysis of longitudinal clinical trials with missing data using reference-based imputation. Journal of Biopharmaceutical Statistics. 2016;26:924–36.
3. Floden L, Bell ML. Imputation strategies when a continuous outcome is to be dichotomized for responder analysis: a simulation study. BMC Medical Research Methodology. 2019;19:161.
4. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test (Madr). 2009;18:1–43.
5. U.S. Food and Drug Administration. FDA briefing document. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/202293Orig1s000MedR.pdf [cited on 11 September 2019].
6. Lehmann EL, Casella G. Theory of point estimation. 2nd ed. New York: Springer; 1998.
7. National Research Council (US) Panel on Handling Missing Data in Clinical Trials. The Prevention and Treatment of Missing Data in Clinical Trials [Internet]. Washington (DC): National Academies Press (US); 2010 [cited 2019 Oct 14]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK209904/
8. Little RJ, D'agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. New England Journal of Medicine. 2012;367:1355–1360.
9. ICH E9(R1) EWG. Final Concept Paper E9(R1): Addendum to Statistical Principles for Clinical Trials on Choosing Appropriate Estimands and Defining Sensitivity Analyses in Clinical Trials, Oct. 2014 [Internet]. 2014 [cited 2018 Jan 17]. Available from: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9__R1__Final_Concept_Paper_October_23_2014.pdf
10. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Step 2b. [Internet]. 2017 [cited 2017 Sep 8]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/08/WC500233916.pdf
11. ICH Harmonised Tripartite Guideline. Note for Guidance on Statistical Principles for Clinical Trials, ICH Topic E9 [Internet]. 1998 [cited 2017 Sep 1]. Available from: http://home.att.ne.jp/red/akihiro/e9step4.doc
12. ICH Official web site: ICH [Internet]. [cited 2019 Oct 14]. Available from: https://www.ich.org/
13. ICH Harmonised Guideline Addendum On Estimands And Sensitivity Analysis In Clinical Trials To The Guideline On Statistical Principles For Clinical Trials E9(R1). ICH Consensus Guideline. [Internet]. 2019 [cited 2019 Dec 4]. Available from: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf
14. Fletcher C, Tsuchiya S, Mehrotra DV. Current Practices in Choosing Estimands and Sensitivity Analyses in Clinical Trials: Results of the ICH E9 Survey. Drug Inf J. 2017;51:69–76.
15. European Medicines Agency [Internet]. European Medicines Agency. [cited 2019 Oct 14]. Available from: https://www.ema.europa.eu/en
16. Clinical efficacy and safety guidelines [Internet]. European Medicines Agency. 2018 [cited 2019 Oct 14]. Available from: https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-guidelines/clinical-efficacy-safety-guidelines
17. Marketing authorisation [Internet]. European Medicines Agency. 2018 [cited 2019 Oct 14]. Available from: https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation
18. Committee for Medicinal Products Human Use (CHMP) [Internet]. European Medicines Agency. 2018 [cited 2019 Oct 14]. Available from: https://www.ema.europa.eu/en/committees/committee-medicinal-products-human-use-chmp
19. European Medicines Agency pre-authorisation procedural advice for users of the centralised procedure. [Internet]. [cited 2019 Dec 2]. Available from: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/european-medicines-agency-pre-authorisation-procedural-advice-users-centralised-procedure_en-0.pdf
20. Scientific advice and protocol assistance [Internet]. European Medicines Agency. 2018 [cited 2019 Nov 6]. Available from: https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-advice-protocol-assistance
21. ICH M11: Clinical Electronic Structured Harmonised Protocol (CeSHarP) [Internet]. Available from: https://www.ich.org/page/multidisciplinary-guidelines

22. Keene ON, Ruberg S, Schacht A, Akacha M, Lawrance R, Berglind A, et al. What matters most? Different stakeholder perspectives on estimands for an invented case study in COPD. Pharm Stat. 2020;

23. Oude Rengerink K, Mitroiu M, Teerenstra S, Pétavy F, Roes KCB. Rethinking the intention-to-treat principle: one size does not fit all. Journal of Clinical Epidemiology [Internet]. 2020 [cited 2020 May 9]; Available from: http://www.sciencedirect.com/science/article/pii/S0895435620301578

Appendix 1. Structure of information pertaining to estimand attributes and their description

Attribute	Type of document (Source)	Variable/ Endpoint	Population	Population-level summary	Analysis/ imputation method	Intercurrent events and strategies to account for them
Relevant section from corresponding source used for data extraction	Guideline	· Entire guideline content	· Entire guideline content	· Entire guideline content	· Entire guideline content	· Entire guideline content
	Sponsor documentation	· Study objectives · Primary objective · Study design · Primary efficacy · Statistical Methods · Supportive Analyses on the Primary Efficacy Endpoint	· Study objectives · Primary objective · Study design · Inclusion/exclusion criteria · Primary efficacy · Statistical Methods · Supportive Analyses on the Primary Efficacy Endpoint · Study subjects	· Study objectives · Primary objective · Study design · Primary efficacy · Statistical Methods · Supportive Analyses on the Primary Efficacy Endpoint	· Study design · Primary Efficacy · Statistical Methods · Supportive Analyses on the Primary Efficacy Endpoint	· Prior and Concomitant Therapy and Diseases · Prohibited Therapies · Rescue Medication · Prior and Concomitant Diseases · Contraindications · Precautions · Treatment Compliance · Termination of Study · Primary Efficacy · Statistical Methods · Study subjects · Supportive Analyses on the Primary Efficacy Endpoint · Protocol Violations and Deviations · Discontinuations (tables) · Results (tables)
	Regulatory questions	· All clinical efficacy MOs and OCs*	· All clinical efficacy MOs and OCs	· All clinical efficacy MOs and OCs	· All clinical efficacy MOs and OCs	· All clinical efficacy MOs and OCs

*Regulatory questions are major objections (MO) and other concerns (OC).

Appendix 2. Data extraction and interpretation form

Summary conclusion: [i.e. Attributes not specified apart from the endpoints (many), of interest are IEs concomitant medication and change in background medication, not specified how to be handled, only suggested that these should be clearly thought of in advance and pre-specified in the analysis.]
[Relevant verbatim extracts from each corresponding type of document, corresponding to estimand attributes, statistical analysis and imputation methods.]
Data collection and handling: []
MOV: []
NAF: [i.e. adverse events, used only in safety analyses.]

PVs: []

Other IEs of interest: []
MOV=missing outcome values, NAF=not accounted for, PV=protocol violation

Data extraction				**Data interpretation**	
Variable/ endpoint	Population	Population-level summary	Analysis/ imputation method	Implied intercurrent events	Implied strategy/ies to account for intercurrent events
[]	[]	[]	[]	[]	[]

Appendix 3. Disease guidelines and corresponding products

No.	Disease guideline	EMA Document number	Product name
1	Pain	CPMP/EWP/252/03 Rev. 1	Prialt
2	Bipolar disorder	CPMP/EWP/567/98	Adasuve
3	Amyotrophic lateral sclerosis	CPMP/EWP/565/98	Rilutek
4	Attention deficit hyperactivity disorder (ADHD)	EMEA/CHMP/EWP/431734/2008	Intuniv
5	Duchenne and Becker muscular dystrophy	EMA/CHMP/236981/2011, Corr. 11	Translarna
6	Multiple sclerosis	CHMP/771815/2011 Rev. 2	Ocrevus
7	Depression	CPMP/EWP/518/97, Rev. 1	Xeristar
8	Epileptic disorders	CHMP/EWP/566/98 Rev.2/Corr	Briviact
9	Parkinson's disease	CHMP/330418/2012 Rev. 2	Ongentys
10	Generalised anxiety disorder	CPMP/EWP/4284/2002	Xeristar
11	Schizophrenia	EMA/CHMP/40072/2010 Rev. 1	Reagila
12	Alcohol dependence	EMA/CHMP/EWP/20097/2008	Selincro
13	Smoking cessation	Doc. Ref. CHMP/EWP/369963/05	Champix
14	Alzheimer's disease and other dementias	Doc. Ref. CPMP/EWP/553/95 Rev. 1	Ebixa
15	Insomnia	Publication available at https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0773.1992.tb00464.x	Circadin
16	Acute stroke	CPMP/EWP/560/1998	Lixiana
17	Chronic heart failure	CPMP/EWP/235/95 Rev. 1	Entresto
18	Hypertension	CPMP/EWP/238/95 Rev. 2	Edarbi
19	Lipid disorders	EMA/CHMP/748108/2013	Praluent
20	Pulmonary arterial hypertension	CHMP/EWP/356954/08	Uptravi
21	Atrial fibrilation	EMA/CHMP/EWP/213056/2010	Brinavess
22	Treatment of VTE	CPMP/EWP/563/98	Lixiana
23	Acute coronary syndrome	CPMP/EWP/570/98	Kengrexal
24	Myocardial infarction	CPMP/EWP/967/01	Efient
25	Coronary Artery Disease - angina pectoris (the stable angina)	CPMP/EWP/234/95 Rev.1	Ranexa
26	Peripheral-arterial occlusive disease	CPMP/EWP/714/98 rev 1	Brilique
27	Asthma	CHMP/EWP/2922/01 Rev.1	Cinqaero
28	COPD	EMA/CHMP/483572/2012	Incruse
29	Cystic fibrosis	Doc. Ref. EMEA/CHMP/EWP/9147/2008-corr*	Orkambi
30	Anticancer appendix 4 - NSCLC	EMA/CHMP/703715/2012 Rev. 2	Tecentriq

31	Anticancer appendix 4 - prostate cancer	EMA/CHMP/703715/2012 Rev. 2	Tookad
32	Anticancer appendix 4- chronic myeloid leukaemia	EMA/CHMP/703715/2012 Rev. 1	Tasigna
33	Anticancer appendix 4- myelodysplastic syndrome	EMA/CHMP/703715/2012 Rev. 1	Vidaza
34	Anticancer appendix 4- hematopoietic stem cell transplantation	EMA/CHMP/703715/2012 Rev. 1	Zalmoxis

CHAPTER 4

Estimation of treatment effects in short-term depression studies. An evaluation based on the ICH E9(R1) estimands framework

Mitroiu M, Teerenstra S, Oude Rengerink K, Pétavy F, Roes KCB. Estimation of treatment effects in short-term depression studies. An evaluation based on the ICH E9(R1) estimands framework. Pharmaceutical Statistics. 2022;1-21. doi:10.1002/pst.2214

Abstract

Estimands aim to incorporate intercurrent events in design, data collection and estimation of treatment effects in clinical trials. Our aim was to understand what estimands may correspond to efficacy analyses commonly employed in clinical trials conducted before publication of ICH E9(R1). We re-analysed six clinical trials evaluating a new anti-depression treatment. We selected the following analysis methods—ANCOVA on complete cases, following last observation carried forward (LOCF) imputation and following multiple imputation; mixed-models for repeated measurements without imputation (MMRM), MMRM following LOCF imputation and following jump-to-reference imputation; and pattern-mixture mixed models. We included a principal stratum analysis based on the predicted subset of the study population who would not discontinue due to adverse events or lack of efficacy. We translated each analysis into the implicitly targeted estimand, and formulated corresponding clinical questions. We could map six estimands to analysis methods. The same analysis method could be mapped to more than one estimand. The major difference between estimands was the strategy for intercurrent events, with other attributes mostly the same across mapped estimands. The quantitative differences in MADRS10 population-level summaries between the estimands were 4–8 points. Not all six estimands had a clinically meaningful interpretation. Only a few analyses would target the same estimand, hence only few could be used as sensitivity analyses. The fact that an analysis could estimate different estimands emphasises the importance of prospectively defining the estimands targeting the primary objective of a trial. The fact that an estimand can be targeted by different analyses emphasises the importance of prespecifying precisely the estimator for the targeted estimand.

Introduction

The National Research Council (US) Panel on Handling Missing Data in Clinical Trials report on missing data in clinical trials [1] triggered new developments on missing data in academic, regulatory and industry sectors [2–15]. The report and scientific discussions informed the development of the ICH E9(R1) Addendum on estimands and sensitivity analysis in clinical trials, that was recently published [16]. This Addendum aims to enhance the transparency and understanding of treatment effects and how they are estimated by precisely describing a priori the estimands and the associated data collection. It considers that the missing data problem can be better addressed by integrating intercurrent events in the estimation of treatment effects. Missing data are defined in the Addendum as 'Data that would be meaningful for the analysis of a given estimand but were not collected. They should be distinguished from data that do not exist or data that are not considered meaningful because of an intercurrent event'. Intercurrent events are defined as 'Events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest'. The estimand is defined using five attributes: treatment, population, variable, population-level summary and any other strategies for intercurrent events. It is important to understand the potential impact of estimands framework compared to current practice in establishing treatment effects in randomised controlled trials, both scientifically as well as from a regulatory perspective. In this paper we re-analyse six short-term depression trials that supported the initial marketing authorisation of mirtazapine to evaluate common analysis strategies against the new concept of estimands. The main focus is on the impact of dealing with intercurrent events. The treatment attribute defines the regimen involving a precise sequence of interventions. For the depression trials, the investigated treatment is orally administered mirtazapine (or comparator) in addition to standard of care, with dose titrated upward and a pre-defined selection of prohibited concomitant medication. The population attribute describes the target population, in our example 'adults suffering from depression (as defined by DSM IV diagnosis and severity cut-offs at baseline), and not suffering from defined co-morbidities' (we use 'adults suffering from depression' thereafter). The variable attribute describes the clinical outcome to be obtained for each patient at scheduled visits. For most of the six trials, the primary outcome variable was MADRS10 total score to be obtained at baseline, at weeks 1–6 (end of trial). In this paper, the focus is on continuous outcomes. The population-level summary for the variable provides the basis for the comparison between treatment conditions. For example, the difference in mean MADRS10 between the experimental and control arm at a pre-planned timepoint, for example at 6 weeks after initiating treatment.

The Addendum suggests five strategies for addressing intercurrent events: treatment policy, hypothetical, composite variable, while-on-treatment and principal stratum. Treatment policy strategy (actively) ignores the intercurrent event, and uses the outcome irrespective of occurrence of the intercurrent event (enabling the ITT principle as defined in ICH E9 [17]), provided that outcomes can exist after the intercurrent event. If the original measurement of the outcome might not exist after the intercurrent event or might not be meaningful, some of the other suggested strategies (e.g. composite variable) can enable estimation of treatment effects while preserving randomisation, as a way to implement the ITT principle. Hypothetical strategies emulate a scenario where the intercurrent event would not occur, that is for the defined clinical question, the value of the variable that would have been observed without the intercurrent event is of interest. Composite variable strategy incorporates the intercurrent event into the variable definition if the variable is composite (e.g. non-responder imputation), or by assigning a value guided by the reason for missingness and/or its timing in the continuous outcome to reflect the intercurrent event (e.g. assign a worst outcome value from that scale or from values recorded in the control arm). While-on-treatment strategy uses the available outcome up to the last treatment administration, or up to the occurrence of the intercurrent event (e.g. up to rescue medication intake). Principal stratum strategy identifies the population that would have or would have not experienced a certain intercurrent event. There is an interplay between the treatment, population and variable attributes and the strategies for addressing intercurrent events. If a different outcome is chosen, the impact of intercurrent events may not stay the same (or the same event may not be an intercurrent event if it does not affect the outcome anymore). Some of these strategies can be defined at the level of a single attribute (treatment, population or variable), others at the level of a strategy for the remaining intercurrent events [16,18].

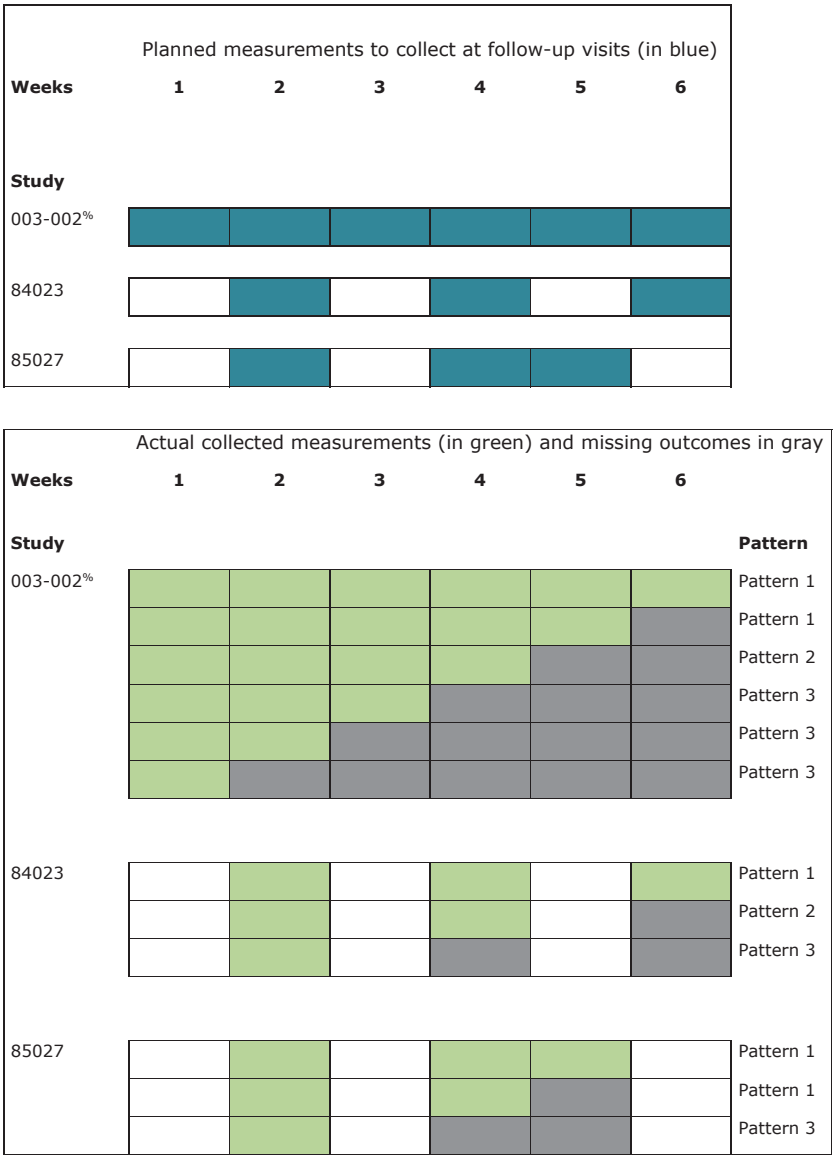
The objective of the present research was to understand what estimands correspond to common efficacy analyses as they were usually applied and are still applied at large, without making distinction between the different intercurrent events. Second, the aim was to assess empirically the impact of choosing between the various analysis methods, by comparing the estimated differences in treatment effect using these methods in short-term depression trials with varying frequencies of the number and type of intercurrent events. In addition, we aimed to explore which methods could be useful as a sensitivity analysis.

Methods

Trial data

We used data from six randomised controlled trials supporting regulatory approval of a new anti-depressive treatment (mirtazapine), designed long before ICH E9(R1) [19]. All trials were double-blind, parallel group; three were placebo-controlled, three were placebo- and active- controlled. Trial treatment duration and follow-up was 5 or 6 weeks, clinical outcome was collected at baseline and at 3 or 6 timepoints post-baseline (Figure 1). The clinical outcome of interest is MADRS10, a widely used score in depression trials, with smaller scores indicating less severe depression [20]. Throughout the trials, patients experienced 'intercurrent events', causing some of the missing outcome values (Supporting Information S1) or impacting values that were observed. However, as these trials were conducted before the ICH E9(R1) era, intercurrent events of potential interest were not registered with enough level of detail.

For illustrative purposes, we treated 'study withdrawal' as approximation of the intercurrent event of interest. Reasons for withdrawal were registered, study withdrawal coincided with stopping treatment and follow-up, and the vast majority of reasons for withdrawal were either occurrence of an adverse event (AE) or experience of lack of efficacy (LoE). We visualised the intercurrent event patterns and outcomes of trials as follows (using study 003-002 as example): the observed and missing data patterns with the heatmaps at trial level (Figure 2) and at arm level (Figure 3), and the longitudinal clinical outcomes, observed and missing, in conjunction with the intercurrent events corresponding to observed patterns (Figure 4). Visualisations for the other five studies are provided in Supporting Information S2. The investigator could record one or more reasons for treatment discontinuation out of the following: adverse events, lack of efficacy, insufficient compliance, efficacy, drug unrelated reasons, and unknown. The reason for discontinuation of study treatment should play a role in determining how the subsequent missing data is handled.



%and for studies 003-020, 003-021 and 003-022

Figure 1. Planned follow-up visits and patterns description. For the top and bottom panel, on the x-axis are displayed the visits number at which measurements were planned to be collected. On the y-axis are three studies with their corresponding design, number of visits and spacing in time. For the bottom panel, the x-axis and y-axis coincide with the description provided above. Additionally, on the right y-axis we displayed possible patterns of observed/ missing outcome data for each distinct trial design.

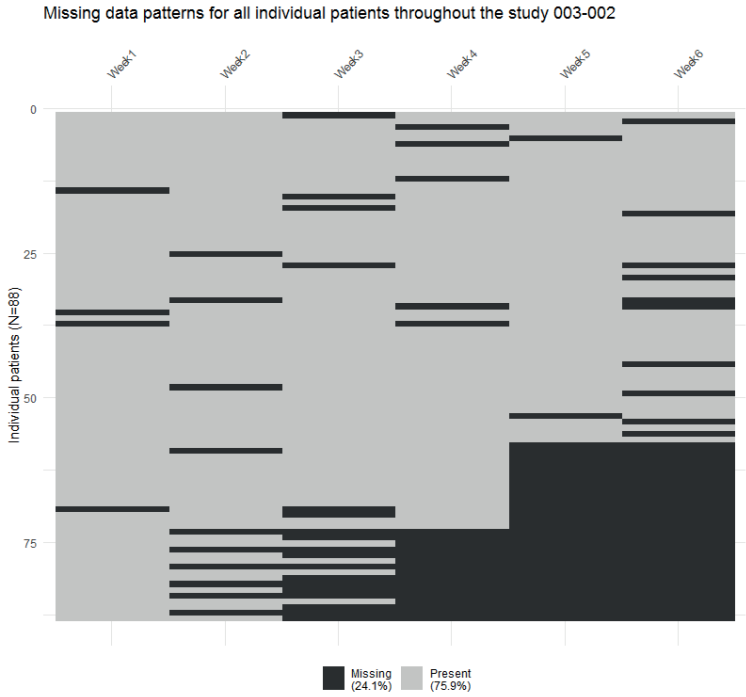


Figure 2. Heatmap missing data patterns for individual patients in study 003-002. This figure displays per visit if data were present or missing for each patient randomised in that particular trial, at trial level.

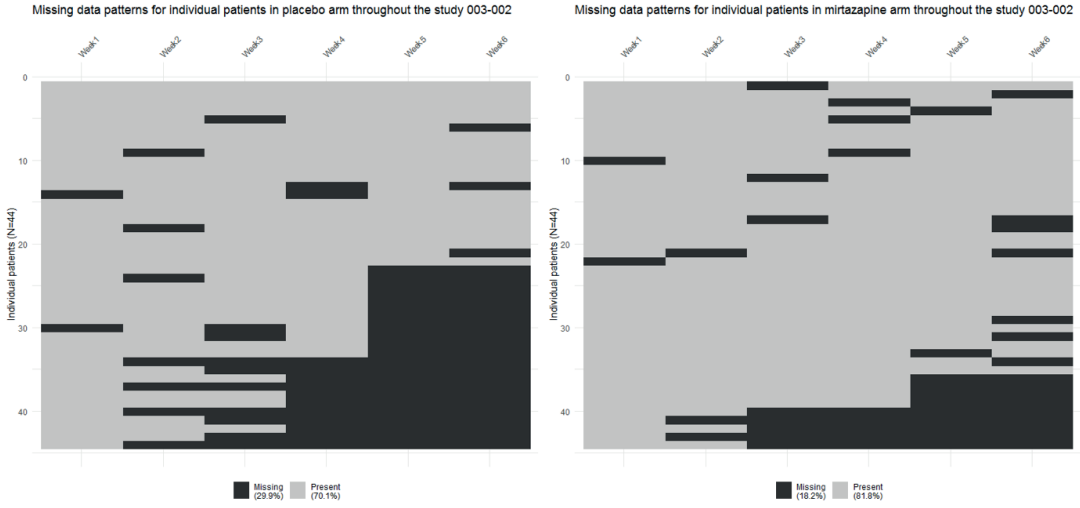


Figure 3. Heatmap missing data at arm level for study 003-002. This figure displays per visit if data were present or missing for each patient randomised in that particular trial, at trial arm level.

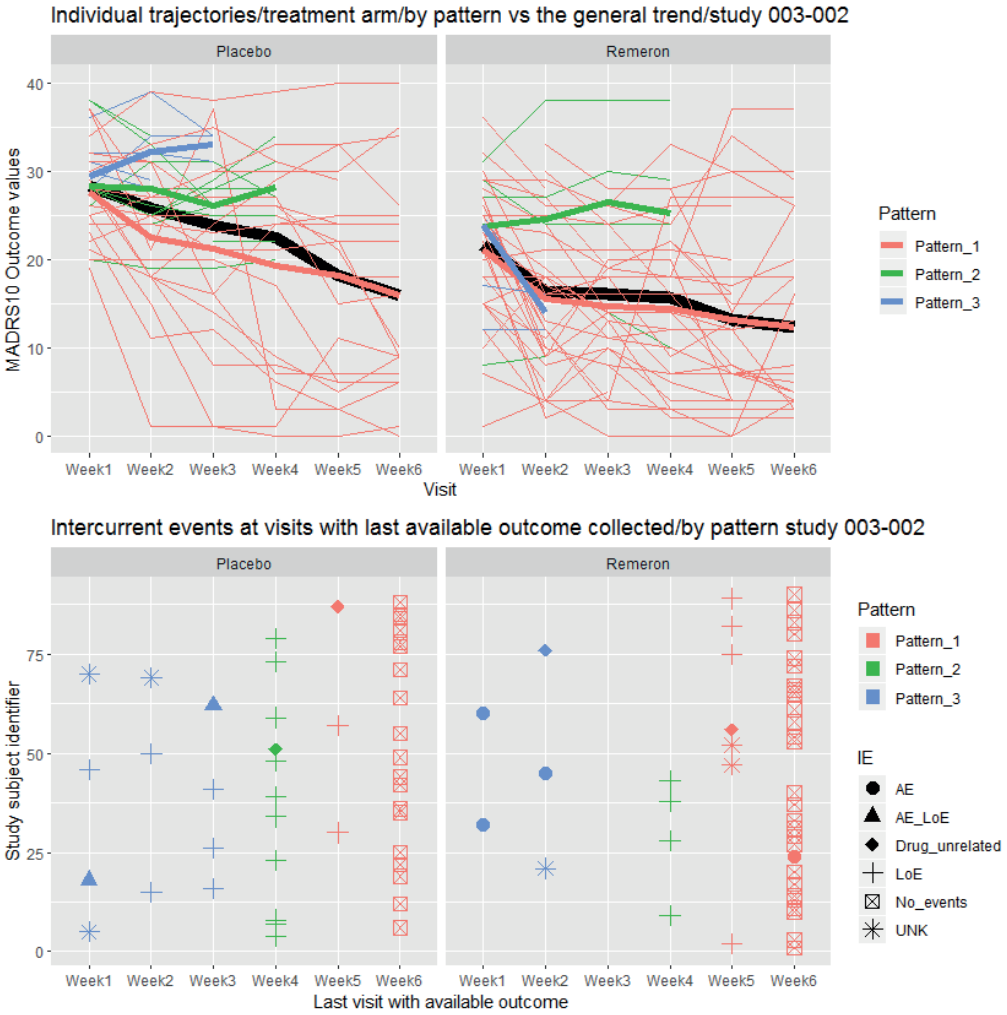


Figure 4. Individual trajectories of patients (Study subject identifier) by treatment arm and pattern of discontinuation with the corresponding intercurrent events. In the top panel each thin line corresponds to a patient and the observed MADRS10 throughout the trial. Each colour corresponds to a different pattern. The thick coloured lines represent the longitudinal group means for each pattern. Each thick line is the average of the thin lines. In the bottom panel the symbols correspond directly and mirror the patients' outcomes from the spaghetti plot (thin lines) from the top panel, and they match 1:1 in colour and timing with the observed outcomes and the pattern they belong to. For instance, we see in the bottom panel in the mirtazapine arm, four green '+'. This symbol ('+') stands for 'lack of efficacy', corresponding to the four thin green lines from the top panel. The colour green corresponds to the late dropouts pattern. The thick green line in the top panel represents the longitudinal group means and trajectory in the pattern of late dropouts. All here dropping out fortuitously due to lack of efficacy. It could be as in the placebo arm, where in the same pattern (thick green line), the patients (thin green lines) dropped out due to 'lack of efficacy' and 'drug-unrelated reasons'. The thick black line corresponds to the group means and longitudinal trajectory of all patients as observed, not differentiated by pattern or intercurrent event.

We used one intercurrent event 'study withdrawal' without making explicit differentiation between the reasons that led to 'study withdrawal', as data collection in older trials did not allow precise enough distinction, nor did the statistical analysis in older trials distinguish between reasons for missing data. Therefore, this choice would not seriously impact the illustration of key differences between analysis models.

Derived estimands and comparison of analysis methods

To understand the treatment effect the methods are estimating, we derived an estimand corresponding to the analysis described, following the estimand structure and definition of the ICH E9(R1) Addendum [16]. The depression trials were used as example for trials with longitudinally observed continuous outcomes. Relative clinical relevance of the derived estimands for depression trials is not specifically evaluated. To quantify and compare the treatment effects estimated using the selected methods, we visualised the population-level summary of treatment effects estimated by each method with 95% CI in Forest plots. We considered a sensitivity analysis any analysis that challenges the assumptions made while targeting the same estimand. All analyses and plots were performed in RStudio software (version 1.2.5042) and relevant packages [21–27].

Selection of methods

We selected common analysis methods for clinical trials (not necessarily targeted to depression trials), based on a survey [28], a review [29] and on the Addendum [16], and included: Analysis of covariance (ANCOVA) [30,31] on complete cases (no missing data in baseline and end of trial outcomes), following last observation carried forward (LOCF) imputation [32], and following multiple imputation [33]; mixed-models for repeated measurements without imputation (MMRM thereafter) [34], MMRM following LOCF imputation and following jump-to-reference (J2R) imputation [35]; pattern-mixture mixed models (PMMM) [36,37], and Principal stratum analysis on a population of interest defined by an intercurrent event (e.g. patients that would not discontinue treatment due to lack of efficacy) via the principal stratification method [18,38,39].

Description and specification of the models

The models are described and specified below (Box 1 and Table 1) for two randomised treatments with six follow-up visits. The ANCOVA estimator uses baseline and end of trial measurements, MMRM uses all repeated measurements.

Results

Studies included

Of the six included studies, 003-002, 003-020, 003-021 and 003-022 were conducted in USA. Studies 84023 and 85027 were conducted in Finland and the United Kingdom, respectively. All studies were multicentre, with parallel groups, patients were randomised to placebo, mirtazapine or amitriptyline in 003-020, 003-021 and 003-022, and to placebo or mirtazapine in 003-002, 84023 and 85027. The sample sizes range between 90 and 150 randomised patients. All studies included adult patients suffering from depression. Study 84023 was an inpatient trial [19]. Further information can be found in the Supporting Information S1.

Missing data and intercurrent events

Data were missing monotonously, intermittently, or both, with different patterns in each study (Figures 1–4, and Supporting Information S2. Percentages of total missing outcome data at trial level ranged from 9.9% (003-022) to 24.1% (003-002 and 003-021), with different distributions between arms and visits. There were mostly monotone missing data in study 84023 (21.3%) and 85027 (12.6%), and both monotone and intermittent missing data in the other studies. (See Figures 2 and 3 and Supporting Information S1.) Percentages of patients with missing outcomes at planned end of study visits differed within studies between treatment groups, and varied between studies, ranging from 16.3% (003-022) to 40.9% (003-021) in mirtazapine arms, from 19.7% (85027) to 56.8% (003-002) in placebo arms, and from 18.4% (003-022) to 31.9% (003-021) in amitriptyline arms.

The type, occurrence, frequency and timing of treatment discontinuation varied between treatment groups within studies, and between studies. Treatment discontinuations were mainly related with AEs or LoE (Figure 4 for study 003-002, Table 3 and Supporting Information S2 for the other five studies).

Box 1 Model descriptions

ANCOVA following multiple imputation of missing MADRS10 outcomes
We used ANCOVA with multiple imputation for missing MADRS10 outcomes as follows [53]: For multiple imputation the predictors used were treatment, baseline MADRS10 outcome values and available outcomes measured at any post-baseline visits. The method used is 'predictive mean matching' [33,53] and we imputed m = 10 datasets. To each imputed dataset, we fitted an ANCOVA model. We modelled the MADRS10 score as a function of treatment and baseline MADRS10 outcome values to derive corresponding sets of statistics (treatment effect estimates and standard errors). We used Rubin's rules for imputation to combine the statistics [54].

Mixed-models for repeated measurements
We modelled post-baseline MADRS10 outcomes with a MMRM [34] model allowing for different visit-specific treatment effects ('saturated'). We modelled the (variation of) repeated measurements by specifying random intercepts (between-patient variation) and an unstructured serial correlation (within-patient variation).

Pattern-mixture mixed models
We modelled a constant difference in treatment effect between patterns and same time profile in the placebo group for each pattern. The choice of patterns is based on timing of intercurrent events that caused monotone missing data. The estimated covariance matrix from the fitted model is used to estimate the weighted standard error for the weighted average. We used three different patterns described below to support the pattern mixture analyses. Monotone missing data refers to all outcome data being missing following an intercurrent event.

1. Completers and quasicompleters: all outcomes available, or last visit outcome available with previous visits outcomes intermittently missing, or only last visit missing outcome with previous visits outcomes available fully or intermittently missing.
2. Late visit dropout: last two visits with monotone missing outcomes, previous visits outcomes available fully or intermittently missing.
3. Early visits dropout: monotone missing outcomes starting from week 2, week 3 or week 4 (missing thereon until the end of trial).

MMRM following jump-to-reference (J2R) multiple imputation
A MMRM model is fitted on the reference arm (placebo) using only the baseline outcome values and time as fixed effects. Missing outcome values following intercurrent events for patients in the experimental arm are imputed in two steps. First, the reference arm model is used to predict the 'fixed part' of the imputed outcome from the baseline outcome and the time of missing value in order to match the patients on the reference arm from which outcomes will be used ('jumped-to'). To the predicted 'fixed' outcome values a random error is added to enable multiple imputation. The added random errors are drawn from the distributions of errors estimated from the MMRM model fit on the reference arm at each corresponding visit j). The final imputed outcome values were not rounded. We then complete the dataset with these imputed values for the experimental arm. Then, we fit a MMRM model on the imputed dataset (on all data, both experimental and reference arm patients) and follow the same steps as for 'MMRM without imputation' analysis method to derive the treatment effect estimate and the standard error.[51]
The intermittent and monotone missing outcomes in the control arm are considered MAR and are not imputed with the jump-to-reference approach. We imputed m = 10 datasets to which we fit the MMRM model to derive m corresponding sets of statistics (treatment effect estimates and standard errors). We used Rubin's rules for imputation to combine these statistics [54].

Principal stratum analysis on a population of interest defined by an intercurrent event (e.g. principal stratum of patients who would not discontinue treatment due to lack of efficacy) – compliers average causal effect (CACE) estimate with respect to the intercurrent event of interest (e.g. treatment discontinuation due to any reason).
A logistic regression model is used to derive propensity scores for each patient (potential outcome) to experience the intercurrent event of interest. A cut-off value is chosen to identify the patients that would not experience the intercurrent event (predicted principal stratum of compliers with regards to treatment discontinuation due to any reason). On this stratum of (imperfectly) identified patients (not the same as the observed compliers or completers), we apply the MMRM model as described above, without any imputation, and with the same model specification. The use of MMRM as an estimator was driven by the longitudinal design with repeated measures; other estimators could also be used, for example ANCOVA.

Estimands corresponding to analysis methods

We were able to map six different estimands to the common analyses investigated in this research (Table 2). The same estimand can be targeted by different analyses, differing through the assumptions made at statistical estimation level. In Table 2 we mention fully only the attributes that are different from those mentioned earlier in the table. If an attribute is the same as in the previously described estimand, it is not explicitly described again. Furthermore, in Table 2, in column 'Description of the other estimand attribute(s) and formulated clinical questions', we provide for all derived estimands the corresponding clinical questions.

Table 1. Model specification and notation

Analysis method	Model	Notation
ANCOVA	$Y_i = \beta_0 + \beta_1 x\ Treatment_i + \beta_2 x\ BaselineValue_i + \varepsilon_i$	Y_i = MADRS10 outcome for patient i at planned end of trial visit at six weeks $\varepsilon_i \sim N\left(0, \sigma^2_{\varepsilon_i}\right)$ $Treatment_i$ = randomised treatment for patient i, indicator is 0 for control and 1 for treatment Average difference in means (δ) between treatment and placebo at end of trial: $\delta = \beta_1$
MMRM	$Y_{ij} = \beta_0 + \beta_1 x\ Treatment_i + \beta_{2F} x\ Time_{ij}^{(F)} + \beta_{3F} x\ Time_{ij}^{(F)} x\ Treatment_i + \beta_4 x\ BaselineValue_i + b_i + \varepsilon_{ij}$	Y_{ij} = MADRS10 outcome for patient i at follow-up visit j j = follow-up visit 1 to 6 $\varepsilon_{ij} \sim N\left(0, \Sigma^2_{\varepsilon_{ij}}\right)$ $b_i \sim N\left(0, D\right)$ $\Sigma^2_{\varepsilon_{ij}}$ = unstructured correlation matrix to model correlation of repeated measurements in time within patient b_i = random intercepts for patients D = covariance matrix of random effects b_i b_i and ε_{ij} are assumed independent $Time_{ij}^{(F)}$ = time index for patient i being at follow-up visit F ($Time_{ij}^{(F)} = 1$ if $j = F$, $Time_{ij}^{(F)} = 0$ if $j \neq F$), $F = 2, \ldots, 6$ β_1 = average difference in MADRS10 between treatment and placebo group at follow-up visit 1 (treatment effect at follow-up visit 1) β_0 = average MADRS10 score in placebo arm at follow-up visit 1 (only post-baseline outcomes are modeled) β_{2F} = average change in MADRS10 for placebo group from first follow-up visit to follow-up visit F β_{3F} = average difference in effect on MADRS10 between follow-up visit F and follow-up visit 1 β_4 = average influence of baseline MADRS10 outcome values on the post-baseline MADRS10 outcomes Average difference in means (δ) between treatment and placebo at visit 6 (at week 6): $\delta_6 = \beta_1 + \beta_{36}$

PMMM	$Y_{ij} = \beta_0 + \beta_1 x\ Treatment_i + \beta_{2F} x\ Time_{ij}^{(F)} + \beta_{3F} x\ Time_{ij}^{(F)} x\ Treatment_i + \beta_4 x\ BaselineValue_i + \beta_{5P} x\ Pattern_i^{(P)} + \beta_{6P} x\ Treatment_i x\ Pattern_i^{(P)} + b_i + \varepsilon_{ij}$	β_{2F} = average change in MADRS10 for placebo group from first follow-up visit to follow-up visit F β_{3F} = average difference between treatment and placebo in the change in MADRS10 between follow-up visit F and follow-up visit 1 β_4 = average influence of baseline MADRS10 score on the post-baseline MADRS10 outcomes P = denotes the pattern for each patient, reference is pattern 1 (completers and quasicompleters) β_{5P} = average difference on MADRS10 in pattern P compared to pattern 1 (completers and quasicompleters) in placebo group at follow-up visit 1 β_{6P} = average difference in effect on MADRS10 between pattern P and pattern 1 (completers and quasicompleters) Difference in means (δ) between treatment and placebo at visit 6 at week 6, taking into account the drop-out patterns: $\delta_6 = \beta_1 + \beta_{36} + \beta_{6P}$ And the estimated treatment effect from the PMMM is the average of these pattern-specific effects weighted by their occurrence in the sample. The estimated covariance matrix from the fitted model is used to estimate the weighted standard error for the weighted average $Adjusted\ \delta_6$.

Table 2. Analysis methods and corresponding derived estimands

No.	Method	Strategy/ies for intercurrent events	Description of the formulated clinical questions and other estimand attribute(s)	Necessary missing data assumptions for estimand
1	ANCOVA on complete cases	Patients with missing outcomes at end of trial because of treatment discontinuation at any timepoint due to AEs/LoE/other reasons are not included in the analysis. Hypothetical strategy for treatment discontinuation due to AEs/LoE/other reasons. Treatment policy strategy for other intercurrent events.	*What is the improvement achieved if no patients stopped the treatment?* *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, had the treatment discontinuation due to AEs/LoE/other reasons not occurred, and if patients who discontinued were to have similar disease course afterwards as patients taking the treatment but did not discontinue, and regardless of other intercurrent events?* **[Estimand 1 ANCOVA]**	MCAR for missing data at end of trial for an unbiased estimate[b]. Assumptions: patients who discontinued treatment due to AEs/LoE/other reasons constitute a random sample of all included patients. No assumption needed for intermittent missing outcomes.
		Patients with missing outcomes at end of trial because of treatment discontinuation at any timepoint due to AEs/LoE/other reasons are not of interest, and are not included in the analysis. Treatment policy strategy for other intercurrent events.	*What is the improvement achieved in those patients that can complete the treatment?* Population: Adults suffering from depression who complete the study and the treatment. **This is not a population that can be defined outside the clinical trial.** Population-level summary: Difference between experimental treatment completers and control[c] treatment completers in mean MADRS10 after six weeks[a] of treatment **There is no estimand corresponding to this commonly employed analysis.** Or this is a principal stratum estimand, but the estimator is biased [55].	If estimation is aimed at the difference between completers in the experimental arm and completers from the control arm, then no assumption is made for missing data.

No.	Method	Strategy/ies for intercurrent events	Description of the formulated clinical questions and other estimand attribute(s)	Necessary missing data assumptions for estimand
2	ANCOVA following LOCF imputation	While on treatment strategy for treatment discontinuation at any timepoint due to AEs/LoE/other reasons. Treatment policy strategy for intercurrent events that lead to intermittent missing outcomes or do not cause missing outcomes but may impact the efficacy estimate.	*What is the improvement achieved before treatment had to be discontinued?* Population-level summary: Difference between experimental treatment and control in mean MADRS10 score **prior to treatment discontinuation** within a maximum of 6 weeks of treatment *What is the treatment effect on MADRS10 in adults suffering from depression prior to treatment discontinuation due to AEs/LoE/other reasons, treatment being administered as the only medication to treat depression, compared to no treatment being taken and regardless of other intercurrent events?* **[Estimand 2 ANCOVA]**	There is no need for additional missing data assumptions. Values following treatment discontinuation are not of interest for the research question. There is no fixed point in time for the contrast, hence no condition to estimate it at a later (or earlier) timepoint for which there would be the need to make assumptions about missing data.
		Treatment policy strategy for treatment discontinuation at any timepoint due to AEs/LoE/other reasons. Treatment policy strategy for the other intercurrent events.	*What is the improvement achieved regardless of any treatment stopping?* *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, regardless of any intercurrent events?* **[Estimand 3 ANCOVA]**	For an unbiased estimate[b] the distribution of the observed outcomes at end of trial and LOCF imputed outcomes at end of trial must be the same as the distribution of the observed and unobserved outcomes at end of trial. Assumptions for imputation: patients who discontinued treatment due to AEs/LoE/other reasons were not to deteriorate or improve their disease course afterwards. No assumption needed for intermittent missing outcomes.

No.	Method	Strategy/ies for intercurrent events	Description of the formulated clinical questions and other estimand attribute(s)	Necessary missing data assumptions for estimand
3	MMRM following LOCF imputation	Treatment policy strategy for treatment discontinuation at any timepoint due to AEs/LoE/other reasons. Treatment policy strategy for the other intercurrent events.	*What is the improvement achieved regardless of any treatment stopping?* *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, regardless of any intercurrent events?* **[Estimand 3 MMRM]**	For an unbiased estimate[§] the distribution of the observed outcomes and imputed outcomes (LOCF) must be the same as the distribution of the observed and unobserved outcomes. Imputed outcomes are the last available outcome values for patients with monotone missing outcomes and are immediately previous available outcome values for patients with intermittent missing outcomes. Assumptions: patients who discontinued treatment due to AEs/LoE/other reasons were not to deteriorate or improve their disease course afterwards.
4	ANCOVA following multiple imputation	Hypothetical strategy for treatment discontinuation at any timepoint due to AEs/LoE/other reasons. Assumptions: adults suffering from depression who discontinued treatment due to AEs/LoE/other reasons were to have similar disease course afterwards as similar patients taking the treatment but did not experience it. Treatment policy strategy for the other intercurrent events.	*What is the improvement achieved if no patients stopped the treatment?* *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, had the treatment discontinuation due to AEs/LoE/other reasons not occurred?* **[Estimand 1 ANCOVA]**	MAR for missing data imputation for an unbiased estimate[b]. Assumptions: patients who discontinued treatment due to AEs/LoE/other reasons were to have similar disease course afterwards as similar patients (based on covariates in the model) who did not discontinue.
		Treatment policy strategy for treatment discontinuation at any timepoint due to AEs/LoE/other reasons. Treatment policy strategy for the other intercurrent events.	*What is the improvement achieved regardless of any treatment stopping?* *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, regardless of any intercurrent events?* **[Estimand 3 ANCOVA]**	MAR for missing data imputation for an unbiased estimate[b].

No.	Method	Strategy/ies for intercurrent events	Description of the formulated clinical questions and other estimand attribute(s)	Necessary missing data assumptions for estimand
5	MMRM without imputation	Hypothetical strategy for treatment interruption or discontinuation at any timepoint due to AEs/LoE/other reasons. Assumptions: adults suffering from depression who discontinued treatment due to AEs/LoE/other reasons were to have similar disease course afterwards as similar patients taking the treatment but did not experience it. Treatment policy strategy for the other intercurrent events.	*What is the improvement achieved if no patients stopped the treatment?* *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, had the treatment discontinuation due to AEs/ LoE/other reasons not occurred?* **[Estimand 1 MMRM]**	MAR for intermittent and monotone missing data for an unbiased estimate[b].
		Treatment policy strategy for treatment discontinuation at any timepoint due to AEs/LoE/other reasons. Treatment policy strategy for the other intercurrent events.	*What is the improvement achieved regardless of any treatment stopping?* *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, regardless of any intercurrent events?* **[Estimand 3 MMRM]**	MAR for intermittent and monotone missing data for an unbiased estimate[b].

No.	Method	Strategy/ies for intercurrent events	Description of the formulated clinical questions and other estimand attribute(s)	Necessary missing data assumptions for estimand
6	MMRM following J2R imputation	A composite strategy for treatment discontinuation at any timepoint due to AEs/LoE/other reasons. Assumptions: adults suffering from depression who discontinued treatment due to AEs/LoE/other reasons were to have similar disease course afterwards as similar patients that were taking other treatment and did not discontinue treatment. Treatment policy strategy for the other intercurrent events. **(!) This interpretation may not correspond to a clinically meaningful question.**	*What is the improvement achieved when some patients stop treatment, but clinical outcome is assigned to no improvement after stopping treatment?* Variable: composite of all measured MADRS10 outcomes for patients who do not discontinue treatment, and, for patients who discontinue: of measured MADRS10 before discontinuation, and of assigned MADRS10 values after discontinuation, that were from similar patients from the reference arm that did not discontinue. *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, in terms of the composite variable, and regardless of other intercurrent events?* **[Estimand 4 MMRM]**	For control (reference) arm: MAR for intermittent and monotone missing data for an unbiased estimate[b]. MAR for intermittent missing data.
		Treatment policy strategy for treatment discontinuation at any timepoint due to AEs/LoE/other reasons. Treatment policy strategy for the other intercurrent events.	*What is the improvement achieved regardless of any treatment stopping?* *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, regardless of any intercurrent events?* **[Estimand 3 MMRM]**	For control (reference) arm: MAR for intermittent and monotone missing data for an unbiased estimate[b]. For the experimental arm: MNAR for monotone missing data where missing intervention data is considered to jump to reference. MAR for intermittent missing data.

No.	Method	Strategy/ies for intercurrent events	Description of the formulated clinical questions and other estimand attribute(s)	Necessary missing data assumptions for estimand
7	Pattern-mixture mixed model (PMMM) Patterns based on treatment discontinuation times and type of missing data	Composite strategy 1 for early treatment discontinuation (from week 2, 3 or 4) due to AEs/LoE/other reasons. Composite strategy 2 for late treatment discontinuation (from week 5) due to AEs/LoE/other reasons. Assumptions for composite strategies are different based on timing of treatment discontinuation. Assumptions: adults suffering from depression who discontinued treatment due to AEs/LoE/other reasons were to have afterwards the same disease course as observed before the treatment discontinuation according to the pattern where they belong to. Treatment policy strategy for the other intercurrent events. **(!) This interpretation may not correspond to a clinically meaningful question.**	*What is the improvement achieved when some patients stop treatment, but improvement is assigned to continue unchanged after stopping treatment?* Variable: composite of all measured MADRS10 outcomes for patients who do not discontinue treatment, and, for patients who discontinue, of measured MADRS10 before discontinuation and of assigned MADRS10 values after discontinuation, that were extrapolated from the treatment effect achieved before discontinuation, differently for early or late discontinuation for patients who discontinue. *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, in terms of the composite variable, and regardless of other intercurrent events?* **[Estimand 5 MMRM]**	MAR for intermittent missing data.
		Treatment policy strategy for treatment discontinuation due to AEs/LoE/other reasons. Treatment policy strategy for the other intercurrent events.	What is the improvement achieved regardless of any treatment stopping? *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, regardless of any intercurrent events?* **[Estimand 3 MMRM]**	MNAR for monotone missing data with missing data dependent on the patterns (of missing data). MAR for intermittent missing data.

No.	Method	Strategy/ies for intercurrent events	Description of the formulated clinical questions and other estimand attribute(s)	Necessary missing data assumptions for estimand
8	Principal stratum analysis	Principal stratum strategy for a specific intercurrent event of interest (or multiple intercurrent events aggregated) such as treatment discontinuation at any timepoint due to e.g. AEs). Hypothetical strategy for treatment discontinuation due to any other reason than the one of interest, e.g. due to LoE. Treatment policy for the other intercurrent events.	*What is the improvement achieved in those patients that can/could complete the treatment?* Population: Adults suffering from depression that would not discontinue treatment due to AEs/LoE/other reason; patients that would not discontinue treatment due to the intercurrent event of interest *What is the treatment effect on MADRS10 in adults suffering from depression after six weeks[a] of treatment administered as the only medication to treat depression compared to no treatment being taken, that would not experience the intercurrent event of interest, and regardless of other intercurrent events?* **[Estimand 6 MMRM]**	Data missing in a population of interest is MNAR. The population of interest that did not have missing data in the experimental arm would not have had missing data in the control arm either. Data missing due to other reasons than in the population of interest is MAR.

Abbreviations: AEs, adverse events; LoE, lack of efficacy; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.
[a] End of trial is at 6 weeks for studies 003-002, 84023 and 003-020, 003-021 and 003-022, and at 5 weeks for study 85027 (see Figure 1).
[b] Unbiased estimate refers to the estimate if the outcomes would have been observed fully (no missing outcomes).
[c] Experimental treatment is mirtazapine in all studies, control is placebo in studies 003-002, 84023 and 85027, control is placebo and amitriptyline is active control in studies 003-020, 003-021 and 003-022.

Table 3. Intercurrent events (Treatment discontinuations due to different reasons)

Study	Intercurrent events n(%) Arm	AE	AE & LoE	Drug unrelated	LoE	Unk	Efficacy	Insufficient compliance	Insufficient compliance & LoE	AE & Insufficient compliance	AE & Insufficient Compliance & LoE	Insufficient compliance& Drug unrelated
003-002 (N=88)												
	Placebo	0(0)	2(2.3)	2(2.3)	18(20.5)	3(3.4)	0	0	0	0	0	0
	Mirtazapine	4(4.5)	0(0)	2(2.3)	8(9.1)	3(3.4)	0	0	0	0	0	0
84023 (N=105)												
	Placebo	0(0)	0	4(3.8)	14(13.4)	1(1.0)	0(0)	2(1.9)	1(1.0)	0	0	0
	Mirtazapine	1(1.0)	0	2(1.9)	10(9.5)	3(2.9)	2(1.9)	1(1.0)	1(1.0)	0	0	0
85027 (N=124)												
	Placebo	1(0.8)	0	0(0)	7(5.6)	2(1.6)	0(0)	2(1.6)	0	0(0)	0(0)	1(0.8)
	Mirtazapine	3(2.4)	0	1(0.8)	6(4.8)	3(2.4)	1(0.8)	0(0)	0	1(0.8)	1(0.8)	0(0)
003-020 (N=114)												
	Amitriptyline	4(3.5)	4(3.5)	3(2.6)	0(0)	0(0)	1(0.9)	0	0	0	0	0
	Placebo	2(1.8)	1(0.9)	4(3.5)	3(2.6)	2(1.8)	0(0)	0	0	0	0	0
	Mirtazapine	2(1.8)	2(1.8)	5(4.4)	2(1.8)	2(1.8)	1(0.9)	0	0	0	0	0
003-021 (N=139)												
	Amitriptyline	5(3.6)	3(2.2)	3(2.2)	0(0)	4(2.9)	0	0	0	0	0	0
	Placebo	1(0.7)	2(1.4)	6(4.3)	15(10.8)	4(2.9)	0	0	0	0	0	0
	Mirtazapine	1(0.7)	0(0.0)	3(2.2)	13(9.4)	2(1.4)	0	0	0	0	0	0
003-022 (N=148)												
	Amitriptyline	3(2.0)	1(0.7)	2(1.4)	3(2.0)	0(0)	0(0)	0	0	0	0	0
	Placebo	0(0)	2(1.4)	3(2.0)	8(5.4)	0(0)	1(0.7)	0	0	0	0	0
	Mirtazapine	3(2.0)	0(0)	1(0.7)	3(2.0)	1(0.7)	0(0)	0	0	0	0	0

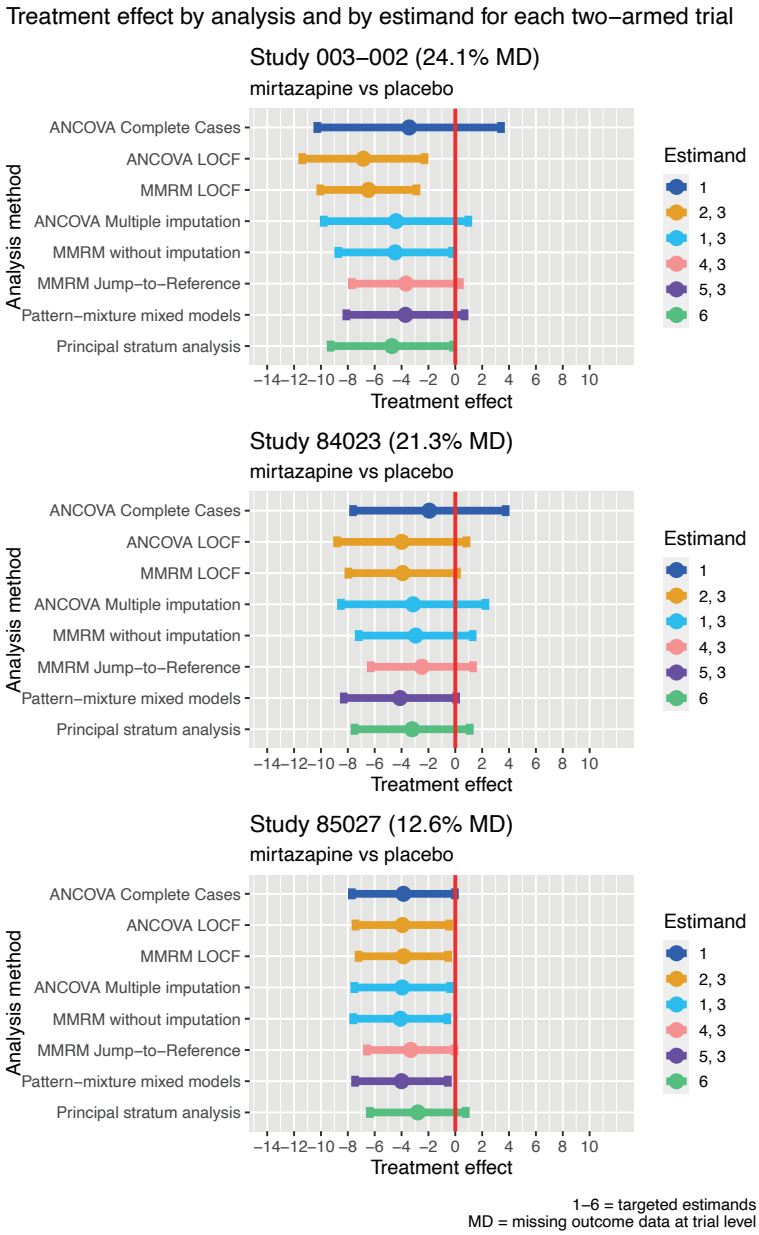


Figure 5. Forest plot treatment effects studies 003-002, 84023 and 85027

The treatment is the same for all estimands (Experimental treatment or control[c] administered as the only medication to treat depression for 6 weeks[a]—see Figure 1). The target population is the same for all estimands (Adults suffering from depression), except for the estimand involving a principal stratum strategy (Adults suffering from depression that would not experience the intercurrent event of treatment discontinuation due to any reason). The variable of interest is the same for all estimands (MADRS10) (See Table 2). The population-level summary is the same for all estimands (Difference between experimental treatment and control in mean MADRS10 score after 6 weeks[a] of treatment, and in view of the variable definition), except for the estimand involving a while-on-treatment strategy (Difference between experimental treatment and control in mean MADRS10 score prior to treatment discontinuation within a maximum of 6 weeks[a] of treatment). Hence, the major differences concern the different strategies for addressing intercurrent events of interest) Table 3.

ANCOVA on complete cases can only be mapped to an estimand after 6 weeks[a] of treatment for the target population under the assumption that completers are a random sample from all patients included in the study. If this were to hold, it would constitute a treatment policy strategy for the intercurrent event.

For analyses that use LOCF imputation (ANCOVA and MMRM), we were able to formulate two estimands that capture the treatment effect in the target population. One estimand could define a treatment effect with a treatment policy

strategy for intercurrent events under the assumption that patients' outcomes remain unchanged after their last observation before stopping treatment. This is not hypothetical in the sense of 'the intercurrent event would not occur'.

The second estimand strategy for intercurrent events we identified for ANCOVA following LOCF imputation is while-on-treatment: the last available assessment is analysed as last value on treatment (what the treatment was able to achieve before it was stopped), and not defined in terms of time since start of treatment.

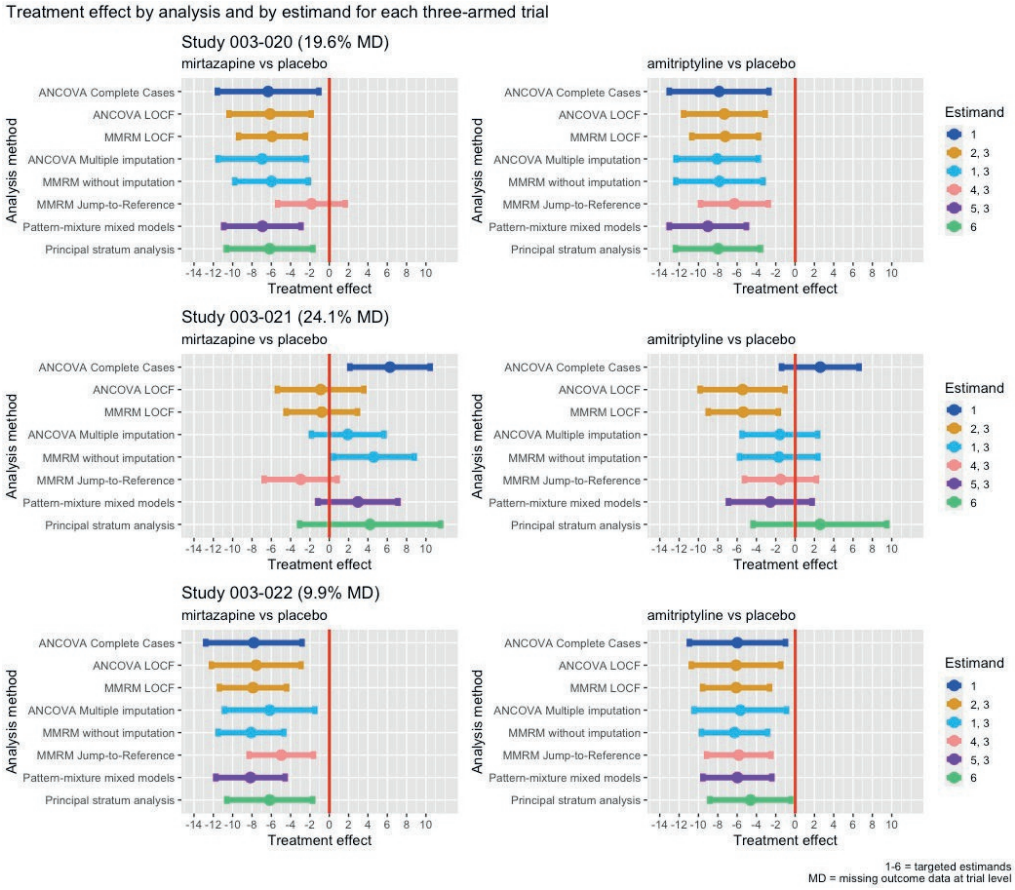


Figure 6. Forest plot treatment effects studies 003-020, 003-021 and 003-022

ANCOVA following multiple imputation and MMRM without imputation essentially target the same estimands under similar assumptions, only following different analysis strategies. They lead to estimands prespecified with a single hypothetical strategy for addressing all intercurrent events, and could only be considered as targeting an estimand with a single treatment policy strategy for addressing all intercurrent events, if the 'missing at random' assumption holds for the occurrence of intercurrent events.

MMRM following J2R imputation and PMMM may aim at an estimand prespecified with a treatment policy strategy for addressing all intercurrent events. They may also aim at a composite variable strategy to ensure an appropriate outcome after the intercurrent event occurred can be included. When the treatment policy strategy is intended, MMRM following J2R imputation makes the strong assumption that following treatment discon-

tinuation, the patient will not take any other treatment than the reference (placebo in these trials). However, depending on the reference treatment this situation might rarely lead to a relevant estimand, unless the reference treatment would be the usual treatment to switch to in case of treatment failure.

In the principal stratum analysis, an attempt is made for an estimand prespecified with a principal stratum strategy for the intercurrent event of interest (and hypothetical strategy for other intercurrent events).

All analyses involving imputation could also be interpreted as trying to estimate a treatment effect having observed all outcomes at end of treatment period, hence, an estimand prespecified with treatment policy as single strategy for addressing all intercurrent events. In absence of actually observed outcomes, an imputation of the missing outcomes completes the dataset and artificially allows a treatment policy strategy. This approach can only lead to a viable estimand prespecified with a single treatment policy strategy for all intercurrent events, if the model and the strong assumptions for imputation match a realistic scenario of the changes to outcomes after the intercurrent event.

Observed differences in effect estimates

Re-analysis of the six trials using the methods described above largely yielded comparable direction of estimated treatment effects across studies. The range of point estimates was of 4–8 points average reduction in MADRS10 in favour of mirtazapine and amitriptyline compared to placebo (Figures 5 and 6, and Supporting Information S3. A clear exception was study 003-021, showing larger differences between analyses (direction and size) and also differences compared to other studies. Additionally, some analyses (ANCOVA on complete cases, MMRM following J2R and PMMM in 003-002; ANCOVA on complete cases, ANCOVA following multiple imputation, MMRM following J2R and principal stratum analysis in 84023; MMRM following J2R in 003-020) deviated from the general correspondence of effect estimates.

Studies 85027 and 003-022 recorded the smallest percentages of treatment discontinuations (12.6% and 9.9%). In 85027, they were mostly due to LoE, balanced between arms. In 003-022, they were mostly due to AEs in amitriptyline and mirtazapine arms, and LoE in placebo arm. There were small or no amounts of intermittent missing data. The direction and size of treatment effects are similar across analyses, principal stratum analysis in study 85027 being slightly different.

Study 003-020 recorded a considerable amount of treatment discontinuations (29.2%), mostly due to AEs or AEs and LoE, and relatively more frequently in the amitriptyline

arm; comparable amount of intermittent missing data as study 003-022. For study 003-020 the direction and size of treatment effects are comparable across analyses, except those yielded by MMRM following jump-to-reference imputation, which are prominently smaller for mirtazapine when compared to placebo. Jump-to-reference imputation assumes that patients who discontinue treatment in the experimental arm continue afterwards with an outcome trajectory as if they did not take treatment (in trial as if randomised to placebo) from treatment discontinuation onwards.

Studies 003-002 and 84023 recorded large amounts of treatment discontinuations (46% and 36%), especially due to LoE in both studies and both arms, with no intermittent missing data in study 84023 and with a large amount of intermittent missing data in study 003-002. Direction and size of treatment effects are comparable across all analyses except those involving LOCF, which showed larger estimates.

Study 003-021 also recorded a large amount of treatment discontinuations (41.3%), especially due to LoE in placebo and mirtazapine arms and especially due to AEs in the amitriptyline arm. Direction and size of treatment effects are heterogeneous across analyses. The completers from the placebo arm recorded better outcomes than both amitriptyline and mirtazapine arms completers, hence the seemingly negative effect suggested by the ANCOVA conducted on complete cases. However, this analysis ignores the different reasons and mechanisms leading to treatment discontinuations between arms. Also, this analysis shows results similar to the principal stratum analysis. The analysis involving jump-to-reference imputation provides contrasting results as treatment discontinuations from treatment arms will be imputed with outcomes from the placebo arm, and since the placebo arm recorded more improvement than the treatment arms, this analysis provides the opposing direction of treatment effects. Although across most studies results were similar, this study demonstrates a larger impact of the choice of estimand.

Overall, MMRM without imputation yielded similar treatment effect estimates as PMMM. When applying PMMM, first an imputation under MNAR is done, and after this step a MMRM analysis is conducted. PMMM seems to result in slightly larger treatment effects in some cases and this could be due to the fact that for each treatment discontinuation pattern, the treatment effect (slope per pattern) is considered to be retained to the end of the trial.

ANCOVA following multiple imputation yields similar treatment effects with MMRM without imputation, with wider confidence intervals generated by ANCOVA as it uses less data (baseline and endpoint outcomes). Supporting Information S3 contains the table with treatment effects and 95% CI reflected in Figures 5 and 6.

Sensitivity analyses

We found different estimators (analyses) targeting the same estimand. When targeting the same estimand, described by the same attributes, these analyses can be sensitivity analysis for each other, provided the assumptions are different and these differences are specified explicitly. For example, when targeting estimand 1 (Table 2), MMRM (MAR) can be a sensitivity analysis for ANCOVA on complete cases (MCAR), but MMRM (MAR) cannot be a sensitivity analysis for ANCOVA following multiple imputation (MAR). When compared to other analyses, if a single attribute is different in an estimand derived from an analysis, then the targeted estimand is different. Consequently, if the targeted estimand is different, that particular analysis cannot be a sensitivity analysis for the other analyses against which is compared. For example, the difference between MMRM following J2R when targeting estimand 4 and PMMM when targeting estimand 5 is at the level of the variable attribute. Hence, they cannot be used as sensitivity analysis.

In our research it is clear that the same derived estimand could correspond to different analyses (e.g. estimand 3 targeted by ANCOVA following LOCF, MMRM following LOCF, ANCOVA following multiple imputation, MMRM, MMRM following J2R and PMMM). Conversely, the same analysis could correspond to different estimands (e.g. PMMM possibly targeting estimand 5 and estimand 3). The choice of analysis intended for sensitivity analysis purposes must match the estimand targeted by the primary analysis.

Discussion

Investigating the analysis methods using the estimands framework demanded to investigate the estimand attributes and the assumptions made at the level of each analysis method at a fine granularity of detail. Estimands highlight there are different intercurrent events (e.g. study discontinuation due to AEs or LoE, instead of generic 'missing data'), different strategies for intercurrent events and different assumptions made by each analysis. For most of re-analysed trials we found quantitatively comparable numerical results, although some were notably different, such as most analyses in study 003-021 or MMRM following J2R in study 003-020. These showed a different size of treatment effect or both a different size and direction of treatment effect when compared to the other studies. These differences were driven by type (mostly LoE), frequency (large amounts of discontinuations due to LoE) and timing of intercurrent events occurrence (early and/or late in the trial).

This mapping exercise was not simple and did not lead to uniquely defined estimands: the same statistical analysis could be matched to more than one estimand; the recip-

rocal is also true. This suggests the need for pre-specification at trial design stage and benefit from the E9(R1) estimands framework as the analysis alone does not clarify what is being estimated.

ANCOVA on complete cases does not unconditionally aim at a meaningful target population, hence may not provide a useful estimand. If the completers represent a random sample from all patients included in the trial, it can be argued whether the treatment policy estimand is estimated. However, because of the strong assumptions needed, the value of this type of analysis is questionable when used in evaluating clinical trials efficacy.

The only interpretable strategy for intercurrent events we identified for analyses involving LOCF, was for ANCOVA following LOCF imputation, and is the while-on-treatment strategy, as in this case study discontinuation equals treatment discontinuation. It is difficult to say in general if a while-on-treatment strategy is relevant or not; it may be of limited value for depression trials of short duration. We could map an estimand with a treatment policy strategy to ANCOVA following LOCF and MMRM following LOCF. However, we think that LOCF imputation was often not applied for a treatment policy strategy per se, but rather to have an evaluable outcome at the end of the trial such that ITT principle could be applied.

We interpreted the composite variable strategy as prespecifying an outcome value to be incorporated in the variable, to reflect the intercurrent event. One well-known application of the composite variable strategy is, for instance, non-responder imputation. If a patient experiences an intercurrent event or withdraws from the study for any reason, then that patient is assigned a value to reflect 'non-response', specifically to reflect the intercurrent event as treatment failure. This could be a value '0' to reflect the intercurrent event as non-response (e.g. if outcome is dichotomous).

Another interpretation of the composite variable strategy could be when using a composite outcome, such as progression-free survival (PFS) for instance. Other interpretations of the composite strategy are also possible. The ICH E9(R1) does not restrict the use to only binary outcomes and suggests that it can be on a continuous scale too. Furthermore, the framework states that 'An intercurrent event is considered in itself to be informative about the patient's outcome and is therefore incorporated into the definition of the variable'. This leaves other options open, and it does not state precisely whether the outcome to be assigned that reflects the intercurrent event should be fixed or stochastic, single general value for all patients or subject-specific. In this research, we interpreted the composite variable strategy also as prespecifying an outcome value

that is non-fixed. Hence, we considered a jump-to-reference imputation and PMMM a composite variable strategy, as the value assigned is non-constant over patients (subject-specific), and it reflects the intercurrent event (e.g. treatment discontinuation).

With some creativity, another strategy for addressing intercurrent events that can be derived from ANCOVA following LOCF and MMRM following LOCF could alternatively be a 'composite variable' strategy: for patients that withdrew from the study, this intercurrent event is incorporated in the variable definition by using the patients' last measured outcome. The Addendum provides a definition that is larger than the definition of a classical categorical composite outcome. The authors' interpretation (for ANCOVA following LOCF) is that it cannot be only one direction ('bad'/'good'). With LOCF, the clinical meaning can go both ways and is difficult to be clinically interpreted; this duplicate meaning is complicated for a composite variable strategy and also depends on the disease (LOCF in Alzheimer's disease vs LOCF in depression). How the intercurrent events are incorporated in the variable definition does matter in order to have an interpretable estimand. Some ways the composite variable strategy is implemented can be of more interest and more plausible than others.

For MMRM without imputation the point estimates and direction were close across studies with those from MMRM following J2R imputation and PMMM, and slightly closer to PMMM results. In the re-analysed six trials, for some of them the treatment effects were small and overall there are differences to be observed—but in most cases the 95% CI were largely overlapping. The observed results in this research suggest the MAR assumption can be robust to deviations, although this is not generally supported [40–42] MMRM could be considered a reliable and relatively simple starting point as primary efficacy analysis for short-term depression trials.

MMRM following J2R imputation can target a treatment policy treatment effect at the end of trial (6 weeks[a]) where outcomes after treatment discontinuation are supposed to follow a stochastic trajectory as if they received reference (placebo in our studies). This assumption needs careful consideration depending on availability of alternative treatments or standard of care in order to avoid over-conservativeness or misinterpretation of treatment effects. Furthermore, depending on the percentages of treatment discontinuations at trial level and on how imbalanced they are between arms, the estimation following J2R imputation aiming at a treatment policy treatment effect at the end of trial can be severely biased. In other disease settings, the 'reference' or the arm which patients 'jump' to may still be a realistic situation. For instance, if the reference is another treatment or standard of care well defined and standardised, and if the 'jump' is part of the standard of care policy.

PMMM can target an estimand prespecified with a single treatment policy strategy for addressing all intercurrent events, where the outcome after treatment discontinuation is assumed to follow a stochastic trajectory conditional on the timing of discontinuation. This assumption is difficult to justify from a clinical viewpoint or verify because very few data are collected after treatment discontinuation. Indeed, the addendum suggests timing can be a differentiating factor. For each of these two intercurrent events a different composite variable strategy can be selected. Depending on the patient behaviour, in other settings more than two intercurrent events could be defined. Consequently, the pattern-mixture used in the analysis should be adapted accordingly.

Although it may be debatable and have some limitations, the J2R and pattern-mixture could correspond to clinically relevant questions.

The principal stratum analysis targets a treatment effect in a stratum of patients that likely will not discontinue treatment. This treatment effect is of great interest. However, it can be a difficult estimand to estimate, because for instance, collected covariates may not be strong predictors of the intercurrent event or there are only few intercurrent events of interest. Although principal stratum strategy is one of the five strategies defined by the Addendum, it is not a commonly employed analysis used in estimation of treatment effects in depression. It could be relevant, and it was used in this paper to illustrate how the strategy could be applied and how the treatment effect in such a principal stratum can be estimated.

Lastly, occurrence of missing data is often encountered in clinical trials, especially in longitudinal studies with multiple planned visits for outcome measurement [29,43]. 'Missing data' is a multi-dimensional concept. There is a vast amount of methods developed to deal with missing covariates or missing outcomes [36,37,44–52]. The amount of missing data should not replace clinical rationale as the driver for the constructed target estimand. Trialists and stakeholders should define the primary objective and the corresponding estimand when designing the trial. Given the estimand and thus the strategies for addressing intercurrent events, missing data should be as much as possible limited by design, data collection and the analysis choice. Data that are nevertheless missing at trial completion are then a problem of estimation and not a problem of targeted estimand. It is up to thorough debate and further evaluation (e.g. with sensitivity analyses) whether the estimator of the targeted estimand is biased and how biased it is and how reliant it is on modelling assumptions regarding missing data not addressed in the estimand.

Our research has several strengths: we systematically translated a range of established analysis methods into estimands according to the E9(R1) Addendum, with a full

description of its attributes and the clinical question of interest. We provide in-depth and transparency-enhancing details regarding embedded and implicit decisions, and assumptions made, at analysis or imputation level for each method. We acknowledge the estimands are not free of assumptions, but being more explicit about the assumptions made, will facilitate the understanding and interpretability of estimands.

Our research has limitations: This research does not start from formulating de novo meaningful research questions and these trials did not have a priori pre-specified estimands as per E9(R1). Derived estimands are for trials conducted before the estimands framework and we have to make do with available collected data long before the estimands framework. For instance, we do not have data for 'other reasons for treatment discontinuation' and we 'constructed' for illustrative purposes a shared intercurrent event for all analyses (any treatment discontinuation, without differentiating by reason where possible), of which we are fully aware it is not fully clinically justified. However, we think this research will improve understanding of the methods used, and improve comparison of analysis on trials that have been initiated before and after introduction of the estimands framework.

Results and conclusions may not be necessarily generalisable to other disease settings or to other type of trials or designs of trials. We conducted the analyses precisely as they are currently being performed, without making differentiation between reasons for treatment discontinuation. It is anticipated that with differentiation in handling different intercurrent events with different strategies, the analyses will possibly yield different results to some extent. However, the possible expected impact would be small for this particular disease and setting; even with (further) differentiation of strategies by intercurrent events, as AEs occurred early in the trials, LoE later or spread-out throughout the entire trial duration, but there were mostly discontinuations due to LoE.

In most cases, there are two or three different estimands mapped for each analysis method. In absence of a precise clinical question articulated beforehand, more than one interpretation of strategies for intercurrent events could be derived, hence more than one estimand. It is therefore not possible to indicate which of the possibly targeted estimands was the actual intended one. This uncertainty does not exist if we methodically move from estimands through design and analysis. In order to avoid ambiguity regarding targeted estimands, it is paramount to start trial planning from the clinical question to be answered linked with the objective of the trial. This question should be precisely mirrored in the estimand attributes, before deciding which is the suitable estimator for it.

Conclusion

In the re-analysed trials, the quantitative differences between the population-level summaries of these estimands were overall small, so in this particular example there was limited impact on the clinical interpretation of the trial results. Not all six estimands had a clinically meaningful interpretation. Only a few analyses would target the same estimand, hence by definition few could be used as sensitivity analyses. The fact that an analysis could estimate different estimands emphasises the importance of prospectively defining the estimands targeting the primary objective of a trial. The fact that an estimand can be targeted by different analyses emphasises the importance of prespecifying precisely the estimator for the targeted estimand.

References

1. National Research Council (US) Panel on Handling Missing Data in Clinical Trials. The Prevention and Treatment of Missing Data in Clinical Trials [Internet]. Washington (DC): National Academies Press (US); 2010 [cited 2019 Oct 14]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK209904/

2. Akacha M, Bretz F, Ruberg S. Estimands in clinical trials - broadening the perspective. Statistics in Medicine. 2017 Jan 15;36(1):5–19.

3. Little R, Kang S. Intention-to-treat analysis with treatment discontinuation and missing data in clinical trials. Statistics in Medicine. 2015 Jul 20;34(16):2381–90.

4. Little RJ, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Neaton JD, et al. The design and conduct of clinical trials to limit missing data. Stat Med. 2012 Dec 10;31(28):3433–43.

5. Permutt T. A taxonomy of estimands for regulatory clinical trials with discontinuations. Statistics in Medicine. 2016 Jul 30;35(17):2865–75.

6. Permutt T. Sensitivity analysis for missing data in regulatory submissions. Statistics in Medicine. 2016;35(17):2876–9.

7. LaVange LM. The Role of Statistics in Regulatory Decision Making. Therapeutic Innovation & Regulatory Science. 2014 Jan;48(1):10–9.

8. ICH E9(R1) EWG. Final Concept Paper E9(R1): Addendum to Statistical Principles for Clinical Trials on Choosing Appropriate Estimands and Defining Sensitivity Analyses in Clinical Trials, Oct. 2014 [Internet]. 2014 [cited 2018 Jan 17]. Available from: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9__R1__Final_Concept_Paper_October_23_2014.pdf

9. ICH E9(R1) EWG. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Step 2b [Internet]. [cited 2017 Sep 8]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/08/WC500233916.pdf

10. Scharfstein DO. A constructive critique of the draft ICH E9 Addendum. Clinical Trials. 2019 Jun 11;1740774519853566.

11. Hernán MA, Scharfstein D. Cautions as Regulators Move to End Exclusive Reliance on Intention to Treat. Annals of Internal Medicine [Internet]. 2018 Mar 20 [cited 2018 Mar 22]; Available from: http://annals.org/article.aspx?doi=10.7326/M17-3354

12. Akacha M. Choosing Measures of Treatment Benefit: Estimands and Beyond. CHANCE. 2019 Oct 2;32(4):12–7.

13. Mallinckrodt CH, Lin Q, Lipkovich I, Molenberghs G. A structured approach to choosing estimands and estimators in longitudinal clinical trials. Pharmaceut Statist. 2012 Nov 1;11(6):456–61.

14. Mallinckrodt CH, Bell J, Liu G, Ratitch B, O'Kelly M, Lipkovich I, et al. Aligning Estimators With Estimands in Clinical Trials: Putting the ICH E9(R1) Guidelines Into Practice. Therapeutic Innovation & Regulatory Science. 2019 Apr 7;216847901983697.

15. Rufibach K. Treatment effect quantification for time-to-event endpoints–Estimands, analysis strategies, and beyond. Pharmaceutical Statistics [Internet]. [cited 2018 Nov 28];0(0). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1917

16. ICH E9(R1) EWG. ICH E9(R1) Estimands and Sensitivity Analysis in Clinical Trials. STEP 4 TECHNICAL DOCUMENT. [Internet]. [cited 2019 Dec 4]. Available from: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf

17. E9 Statistical Principles for Clinical Trials. 1998 Current Step version; Available from: https://database.ich.org/sites/default/files/E9_Guideline.pdf

18. Frangakis CE, Rubin DB. Principal Stratification in Causal Inference. Biometrics. 2002 Mar;58(1):21–9.

19. Drugs@FDA: FDA-Approved Drugs [Internet]. [cited 2020 Apr 1]. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/nda/96/020415Orig1s000rev.pdf

20. Committee for Medicinal Products for Human Use (CHMP). Guideline on clinical investigation of medicinal products in the treatment of depression. EMA/CHMP/185423/2010 Rev. 2 previously (CPMP/EWP/518/97, Rev. 1) [Internet]. 2013 [cited 2020 Apr 1]. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-investigation-medicinal-products-treatment-depression_en.pdf

21. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

22. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2019). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-140, <URL: https://CRAN.R-project.org/package=nlme>.

23. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

24. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686.

25. Thomas Lin Pedersen (2019). patchwork: The Composer of Plots. R package version 1.0.0. https://CRAN.R-project.org/package=patchwork.

26. Richard Iannone, Joe Cheng and Barret Schloerke (2020). gt: Easily Create Presentation-Ready Display Tables. R package version 0.2.0.5. https://CRAN.R-project.org/package=gt.

27. Nicholas Tierney, Di Cook, Miles McBain and Colin Fay (2020). naniar: Data Structures, Summaries, and Visualisations for Missing Data. R package version 0.5.0. https://CRAN.R-project.org/package=naniar.

28. Fletcher C, Tsuchiya S, Mehrotra DV. Current Practices in Choosing Estimands and Sensitivity Analyses in Clinical Trials: Results of the ICH E9 Survey. Drug Inf J. 2017 Jan;51(1):69–76.

29. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. BMC Medical Research Methodology. 2014 Nov 19;14(1):118.

30. Philippas D. Analysis of Covariance (ANCOVA). In: Michalos AC, editor. Encyclopedia of Quality of Life and Well-Being Research [Internet]. Dordrecht: Springer Netherlands; 2014 [cited 2020 Apr 1]. p. 157–61. Available from: https://doi.org/10.1007/978-94-007-0753-5_82

31. Senn S. Change from baseline and analysis of covariance revisited. Statistics in Medicine. 2006;25(24):4334–44.

32. Lachin JM. Fallacies of Last Observation Carried Forward. Clin Trials. 2016 Apr;13(2):161–8.

33. Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation. 2006 Dec;76(12):1049–64.

34. Siddiqui O, Hung HMJ, O'Neill R. MMRM vs. LOCF: A Comprehensive Comparison Based on Simulation Study and 25 NDA Datasets. Journal of Biopharmaceutical Statistics. 2009 Feb 25;19(2):227–46.

35. Cro S, Morris TP, Kenward MG, Carpenter JR. Reference-based sensitivity analysis via multiple imputation for longitudinal trials with protocol deviation. Stata J. 2016 Apr;16(2):443–63.

36. Thijs H, Molenberghs G, Michiels B, Verbeke G, Curran D. Strategies to fit pattern-mixture models. Biostatistics. 2002 Jun 1;3(2):245–65.

37. Molenberghs G, Thijs H, Michiels B, Verbeke G, Kenward MG. Pattern-mixture models. Journal de la société française de statistique. 2004;145(2):49–77.

38. Larsen KG, Josiassen MK. A New Principal Stratum Estimand Investigating the Treatment Effect in Patients Who Would Comply, If Treated With a Specific Treatment. Statistics in Biopharmaceutical Research. 2020 Jan 2;12(1):29–38.

39. Bornkamp B, Rufibach K, Lin J, Liu Y, Mehrotra DV, Roychoudhury S, et al. Principal stratum strategy: Potential role in drug development. Pharmaceutical Statistics. 2021 Feb 23;pst.2104.

40. Mallinckrodt CH, Clark WS, David SR. ACCOUNTING FOR DROPOUT BIAS USING MIXED-EFFECTS MODELS. Journal of Biopharmaceutical Statistics. 2001 Jan 1;11(1–2):9–21.

41. Guizzaro L, Pétavy F, Ristl R, Gallo C. The Use of a Variable Representing Compliance Improves Accuracy of Estimation of the Effect of Treatment Allocation Regardless of Discontinuation in Trials with Incomplete Follow-up. Statistics in Biopharmaceutical Research. 2020 Feb 28;0(0):1–18.

42. Lane P. Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. Pharmaceutical Statistics. 2008;7(2):93–106.

43. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test (Madr). 2009 May 1;18(1):1–43.

44. Rubin DB. Inference and missing data. Biometrika. 1976 Dec 1;63(3):581–92.

45. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology. 2019 Jun 1;110:63–73.

46. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Jun 29;338:b2393.

47. Molenberghs G, Michiels B, Kenward MG, Diggle PJ. Monotone missing data and pattern-mixture models. Statistica Neerlandica. 1998 Jun;52(2):153–61.

48. Little R. Selection and pattern-mixture models. In: Verbeke G, Davidian M, Fitzmaurice G, Molenberghs G, editors. Longitudinal Data Analysis [Internet]. Chapman and Hall/CRC; 2008 [cited 2020 Feb 16]. p. 409–31. Available from: http://www.crcnetbase.com/doi/abs/10.1201/9781420011579.ch18

49. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken, N.J: Wiley; 2002. 381 p. (Wiley series in probability and statistics).

50. Liu GF, Pang L. On analysis of longitudinal clinical trials with missing data using reference-based imputation. Journal of Biopharmaceutical Statistics. 2016 Sep 2;26(5):924–36.

51. Sobel ME, Muthén B. Compliance Mixture Modelling with a Zero-Effect Complier Class and Missing Data. Biometrics. 2012 Dec;68(4):1037–45.

52. Erler NS, Rizopoulos D, Rosmalen J van, Jaddoe VWV, Franco OH, Lesaffre EMEH. Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. Statistics in Medicine. 2016;35(17):2955–74.

53. Buuren S van, Groothuis-Oudshoorn K. Multivariate Imputation by Chained Equations in *R*. Journal of Statistical Software [Internet]. 2011 [cited 2018 May 25];45(3). Available from: http://www.jstatsoft.org/v45/i03/

54. Rubin: Multiple imputation for nonresponse in surveys - Google Scholar [Internet]. [cited 2020 Apr 2]. Available from: https://scholar.google.com/scholar_lookup?title=Multiple+Imputation+for+Nonresponse+in+Surveys&author=DB+Rubin&publication_year=2004&

55. Groenwold RHH, Donders ART, Roes KCB, Harrell FE, Moons KGM. Dealing with missing outcome data in randomized trials and observational studies. Am J Epidemiol. 2012 Feb 1;175(3):210–7.

Supplementary materials

Appendix 1 (Supporting information S1): Descriptive statistics of studies characteristics

1. Study 003-002

Treatment group/N Patients	Randomised	ITT
Mirtazapine	45	44
Placebo	45	44

Visit	collected outcomes (n)	mean MADRS10	SD MADRS10	missing outcomes (ITT-n)	missing %
Mirtazapine					
1	42	21.6	8.1	2	4.5
2	41	16.3	9.1	3	6.8
3	36	16.0	8.3	8	18.2
4	36	15.6	9.8	8	18.2
5	33	13.1	10.1	11	25.0
6	28	12.3	10.6	16	36.4
Placebo					
1	42	28.2	5.3	2	4.5
2	37	25.8	7.9	7	15.9
3	34	23.9	9.2	10	22.7
4	31	22.4	10.8	13	29.5
5	22	18.1	11.5	22	50.0
6	19	15.9	12.0	25	56.8

Clinical outcome parameters based on ITT (All randomised patients with at least 1 post-baseline outcome value available)

Follow-up: Visit number coincides with week number

2. Study 84023

Treatment group/N Patients	Randomised	ITT
Mirtazapine	59	54
Placebo	55	51

Visit	collected outcomes (n)	mean MADRS10	SD MADRS10	missing outcomes (ITT-n)	missing %
Mirtazapine					
1	54	24.9	8.4	0	0.0
2	40	18.1	9.5	14	25.9
3	34	15.2	11.6	20	37.0
Placebo					
1	51	26.4	10.4	0	0.0
2	40	21.6	9.2	11	21.6
3	29	17.4	10.7	22	43.1

Clinical outcome parameters based on ITT (All randomised patients with at least 1 post-baseline outcome value available)

Follow-up: Visit 1 is at week 2, Visit 2 is at week 4, Visit 3 is at week 6

3. Study 85027

Treatment group/N Patients	Randomised	ITT
Mirtazapine	66	63
Placebo	66	61

Visit	collected outcomes (n)	mean MADRS10	SD MADRS10	missing outcomes (ITT-n)	missing %
Mirtazapine					
1	63	20.0	9.4	0	0.0
2	51	12.9	9.2	12	19.0
3	48	11.9	9.5	15	23.8
Placebo					
1	60	20.4	9.3	1	1.6
2	54	16.9	9.8	7	11.5
3	49	14.7	10.3	12	19.7

Clinical outcome parameters based on ITT (All randomised patients with at least 1 post-baseline outcome value available)

Follow-up: Visit 1 is at week 2, Visit 2 is at week 4, Visit 3 is at week 5

4. Study 003-020

Treatment group/N Patients	Randomised	ITT
Amitriptyline	43	38
Mirtazapine	44	39
Placebo	43	37

Visit (n)	collected outcomes	mean MADRS10	SD MADRS10	missing outcomes (ITT-n)	missing %
Amitriptyline					
1	37	25.1	6.7	1	2.6
2	36	20.6	7.8	2	5.3
3	32	19.9	8.3	6	15.8
4	31	16.6	8.5	7	18.4
5	25	15.6	9.1	13	34.2
6	26	13.5	9.5	12	31.6
Mirtazapine					
1	39	23.6	6.9	0	0.0
2	36	21.6	7.4	3	7.7
3	33	18.9	9.4	6	15.4
4	25	17.7	8.4	14	35.9
5	24	16.4	7.4	15	38.5
6	25	14.1	8.1	14	35.9
Placebo					
1	37	26.6	9.9	0	0.0
2	31	25.7	10.5	6	16.2
3	31	25.7	12.1	6	16.2
4	29	23.4	10.9	8	21.6
5	28	20.6	13.3	9	24.3
6	25	21.6	11.7	12	32.4

Clinical outcome parameters based on ITT

Follow-up: Visit number coincides with week number

5. Study 003-021

Treatment group/N Patients	Randomised	ITT
Amitriptyline	50	47
Mirtazapine	50	44
Placebo	50	48

Visit	collected outcomes (n)	mean MADRS10	SD MADRS10	missing outcomes (ITT-n)	missing %
Amitriptyline					
1	45	21.6	7.6	2	4.3
2	41	20.0	7.2	6	12.8
3	34	16.0	8.1	13	27.7
4	36	12.5	7.5	11	23.4
5	31	12.5	8.3	16	34.0
6	32	10.5	7.4	15	31.9
Mirtazapine					
1	42	23.0	7.3	2	4.5
2	40	19.0	9.6	4	9.1
3	39	18.7	10.6	5	11.4
4	39	16.9	11.4	5	11.4
5	26	12.6	8.0	18	40.9
6	26	13.7	8.1	18	40.9
Placebo					
1	48	26.8	7.6	0	0.0
2	42	23.1	9.1	6	12.5
3	36	19.7	10.4	12	25.0
4	32	20.3	10.7	16	33.3
5	23	15.5	10.5	25	52.1
6	21	7.5	5.2	27	56.2

Clinical outcome parameters based on ITT (All randomised patients with at least 1 post-baseline outcome value available)

Follow-up: Visit number coincides with week number

6. Study 003-022

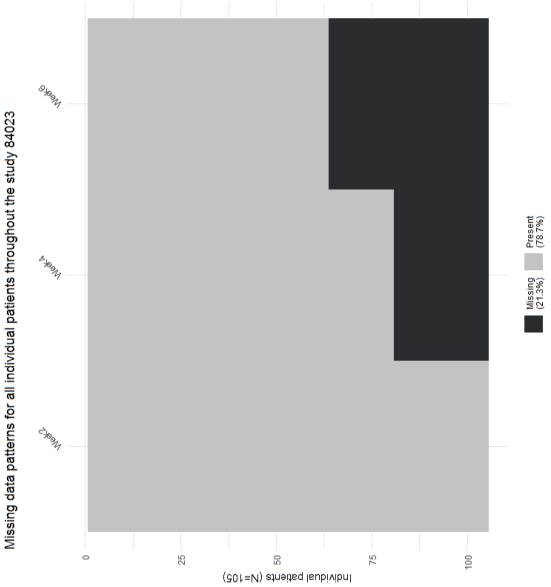
Treatment group/N Patients	Randomised	ITT
Amitriptyline	50	49
Mirtazapine	50	49
Placebo	50	50

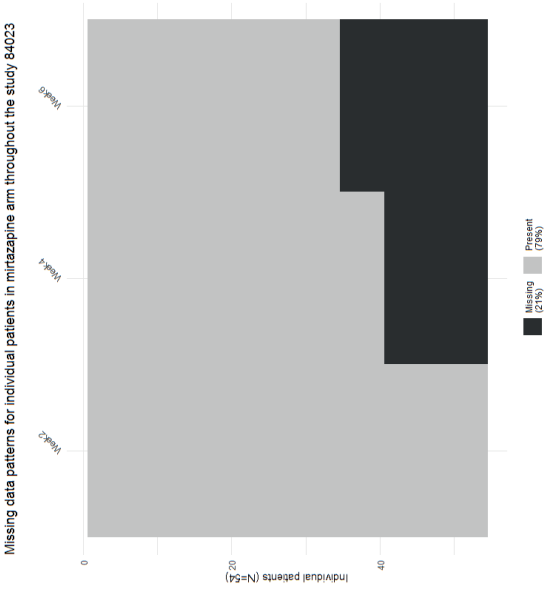
Visit	collected outcomes (n)	mean MADRS10	SD MADRS10	missing outcomes (ITT-n)	missing %
Amitriptyline					
1	49	32.6	6.2	0	0.0
2	47	26.6	6.8	2	4.1
3	45	22.4	7.7	4	8.2
4	42	20.5	9.2	7	14.3
5	40	17.7	9.7	9	18.4
6	40	18.0	10.3	9	18.4
Mirtazapine					
1	49	33.5	6.4	0	0.0
2	47	28.9	7.4	2	4.1
3	46	27.0	7.7	3	6.1
4	43	23.2	8.6	6	12.2
5	42	19.9	9.9	7	14.3
6	41	16.5	12.0	8	16.3
Placebo					
1	50	34.0	5.1	0	0.0
2	48	31.0	5.8	2	4.0
3	46	27.9	8.1	4	8.0
4	45	25.2	9.6	5	10.0
5	42	23.7	10.1	8	16.0
6	38	23.9	10.9	12	24.0

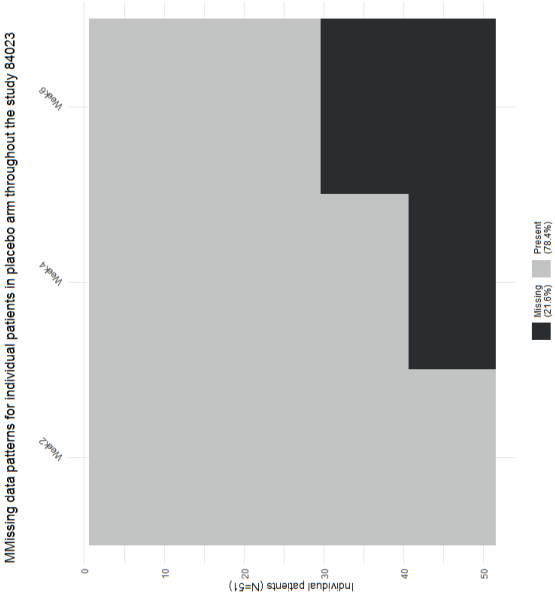
Clinical outcome parameters based on ITT (All randomised patients with at least 1 post-baseline outcome value available)

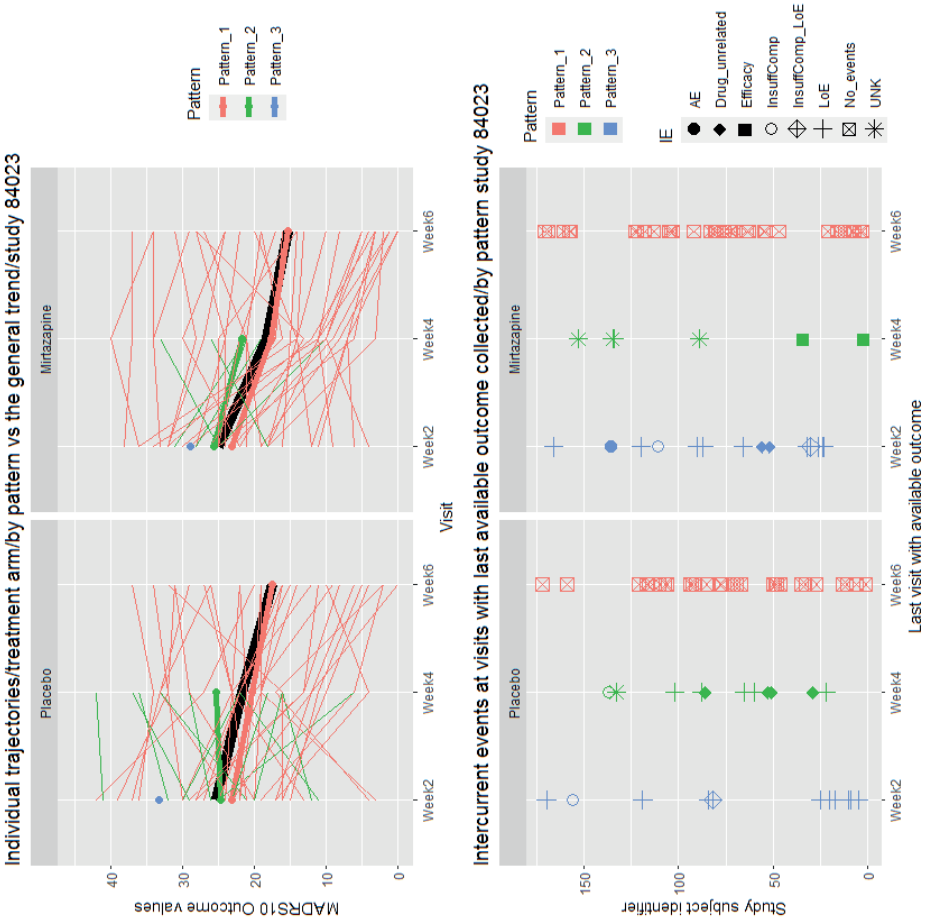
Follow-up: Visit number coincides with week number

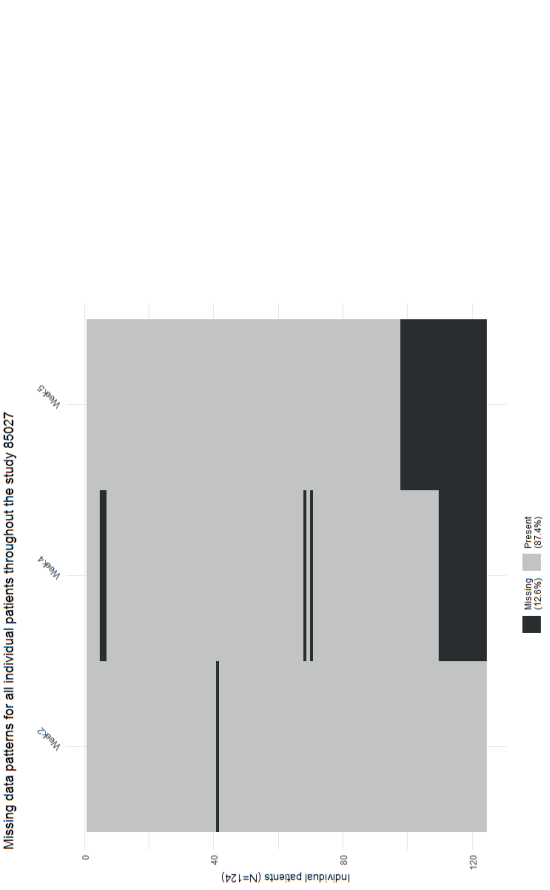
Appendix 2 (Supporting information S2): Heatmaps and individual trajectories for studies 84023, 85027, 003-020, 003-021 and 003-022

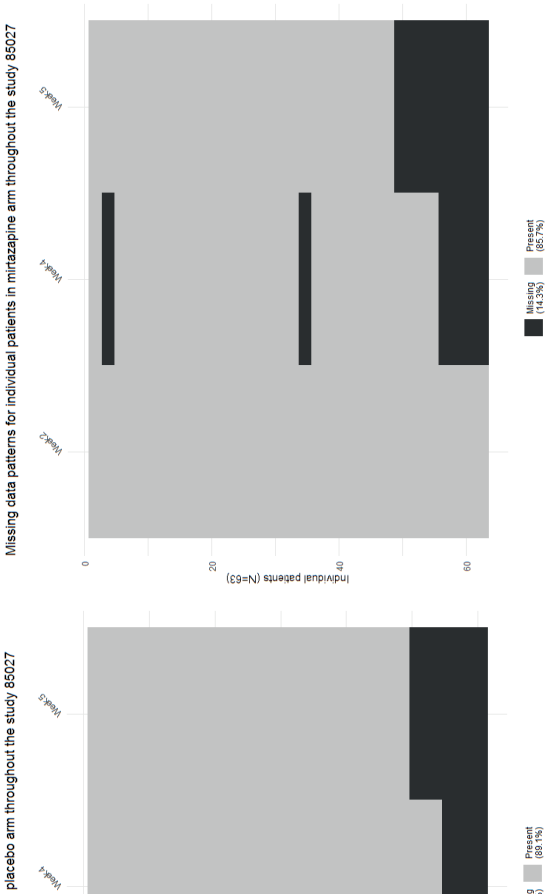


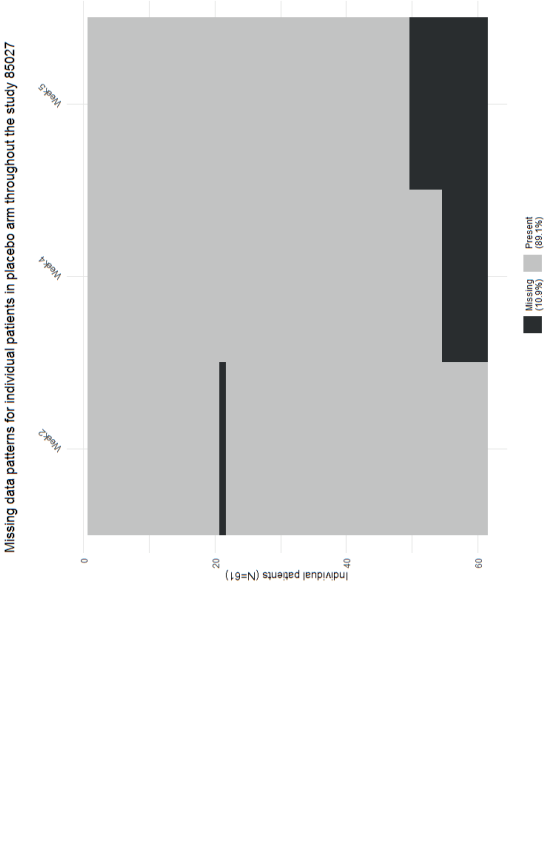


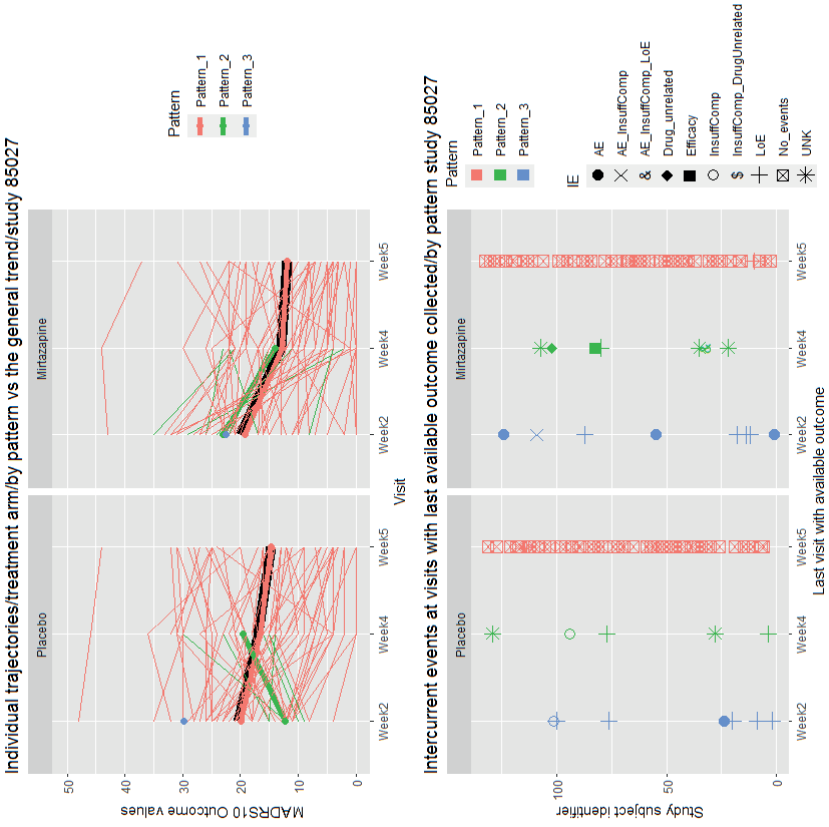


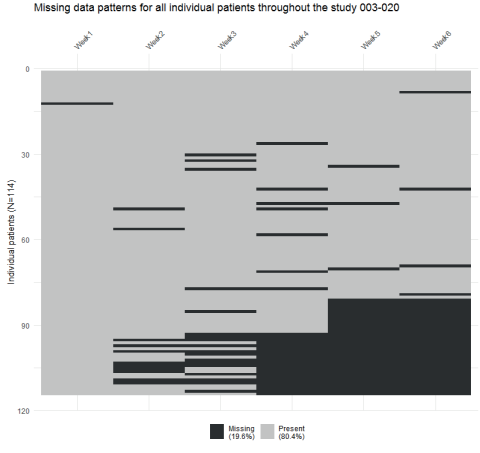

Individual trajectories/treatment arm/by pattern vs the general trend/study 84023
Placebo
Mirtazapine
MADRS10 Outcome values
Visit
Week2
Week4
Week6
Pattern
Pattern_1
Pattern_2
Pattern_3
Intercurrent events at visits with last available outcome collected/by pattern study 84023
Study subject identifier
Last visit with available outcome
IE
AE
Drug_unrelated
Efficacy
InsuffComp
InsuffComp_LoE
LoE
No_events
UNK

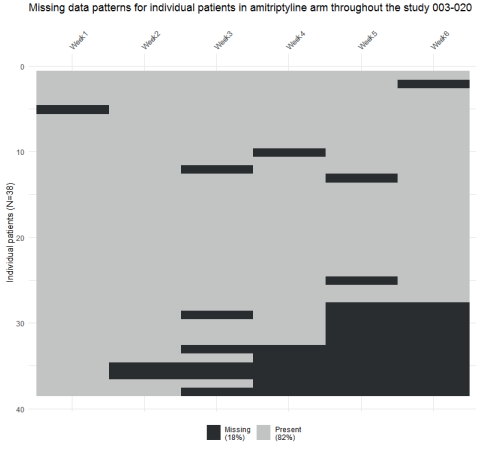

Missing data patterns for all individual patients throughout the study 85027
Week 2
Week 4
Week 5
0
40
80
120
Individual patients (N=124)
Missing (12.6%)
Present (87.4%)

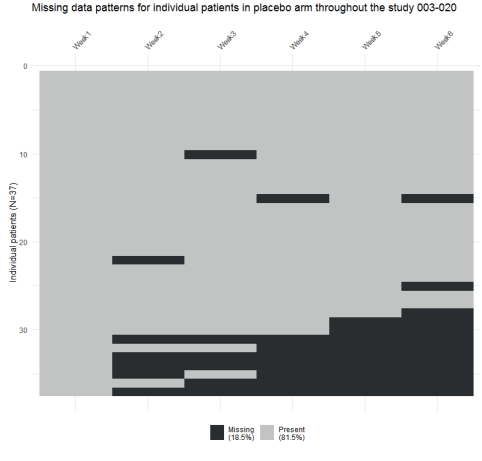

Missing data patterns for individual patients in mirtazapine arm throughout the study 85027
Week 2
Week 4
Week 5
0
20
40
60
Individual patients (N=63)
Missing (14.3%)
Present (85.7%)

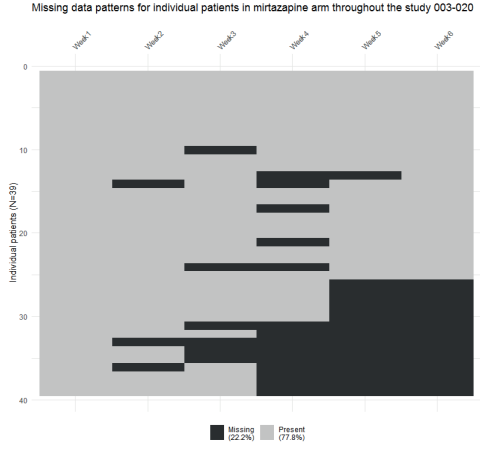

Missing data patterns for individual patients in placebo arm throughout the study 85027
Week 2
Week 4
Week 5
0
20
40
60
Individual patients (N=61)
Missing (10.8%)
Present (89.1%)

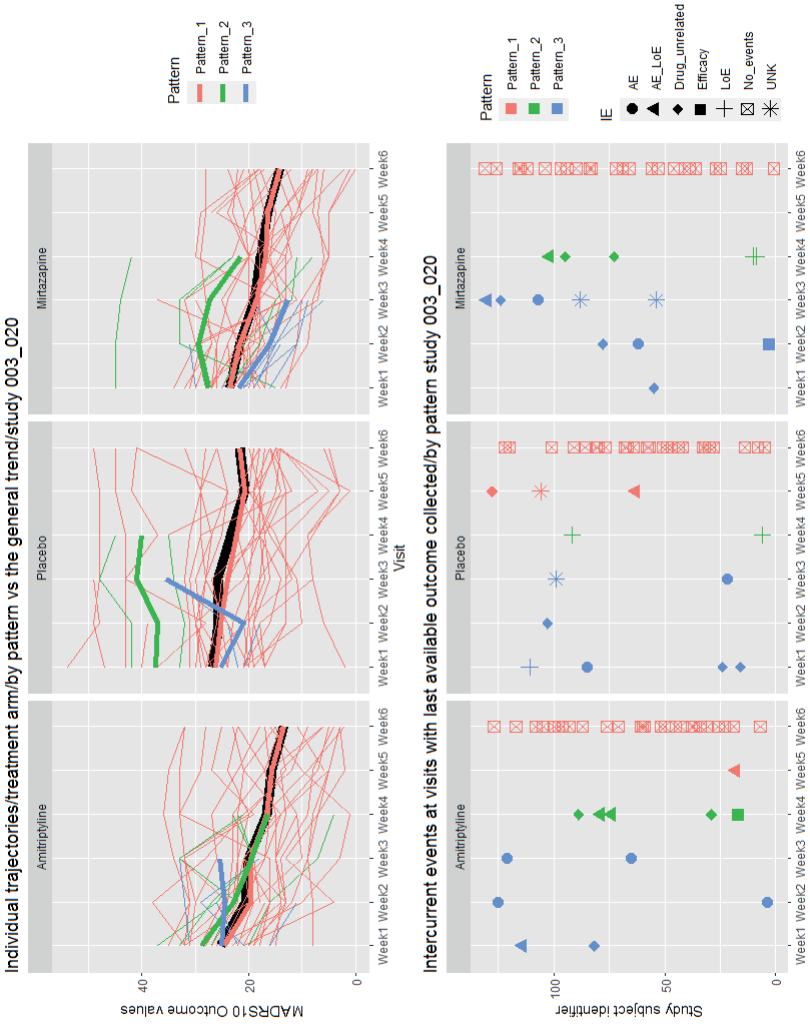

Individual trajectories/treatment arm/by pattern vs the general trend/study 85027
Pattern
Pattern_1
Pattern_2
Pattern_3
Placebo
Mirtazapine
MADRS10 Outcome values
Visit
Week2
Week4
Week5
Intercurrent events at visits with last available outcome collected/by pattern study 85027
IE
AE
AE_InsuffComp
AE_InsuffComp_LoE
Drug_unrelated
Efficacy
InsuffComp
InsuffComp_DrugUnrelated
LoE
No_events
UNK
Study subject identifier
Last visit with available outcome

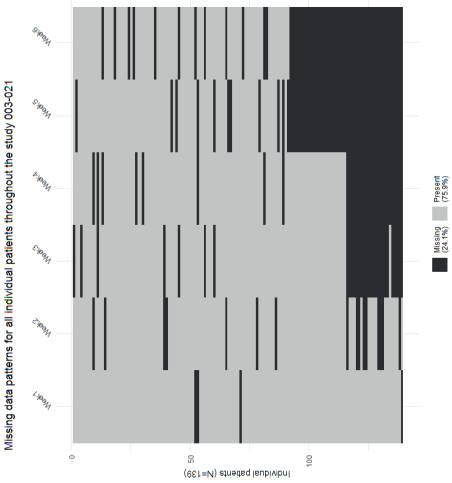

Missing data patterns for all individual patients throughout the study 003-020
Week1
Week2
Week3
Week4
Week5
Week6
0
30
60
90
120
Individual patients (N=114)
Missing (19.5%)
Present (80.4%)

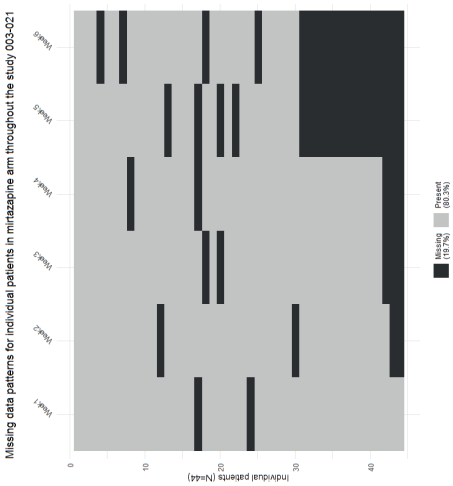

Missing data patterns for individual patients in amitriptyline arm throughout the study 003-020
Week1
Week2
Week3
Week4
Week5
Week6
0
10
20
30
40
Individual patients (N=38)
Missing (18%)
Present (82%)

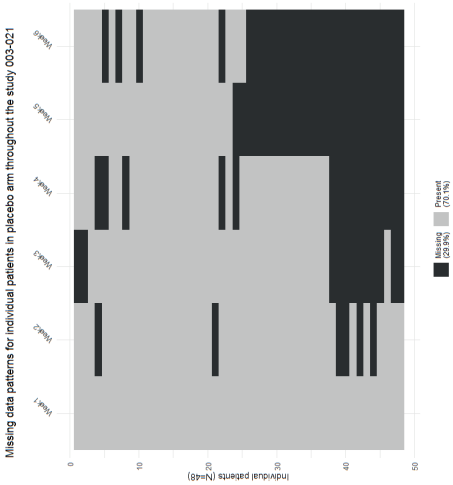

Missing data patterns for individual patients in placebo arm throughout the study 003-020
Week1
Week2
Week3
Week4
Week5
Week6
0
10
20
30
Individual patients (N=37)
Missing (18.5%)
Present (81.5%)

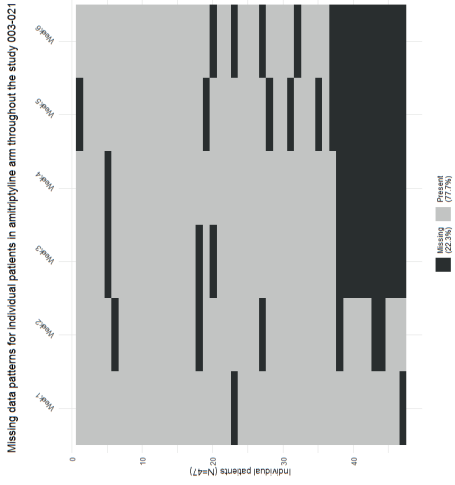

Missing data patterns for individual patients in mirtazapine arm throughout the study 003-020
Week1
Week2
Week3
Week4
Week5
Week6
0
10
20
30
40
Individual patients (N=39)
Missing (22.2%)
Present (77.8%)

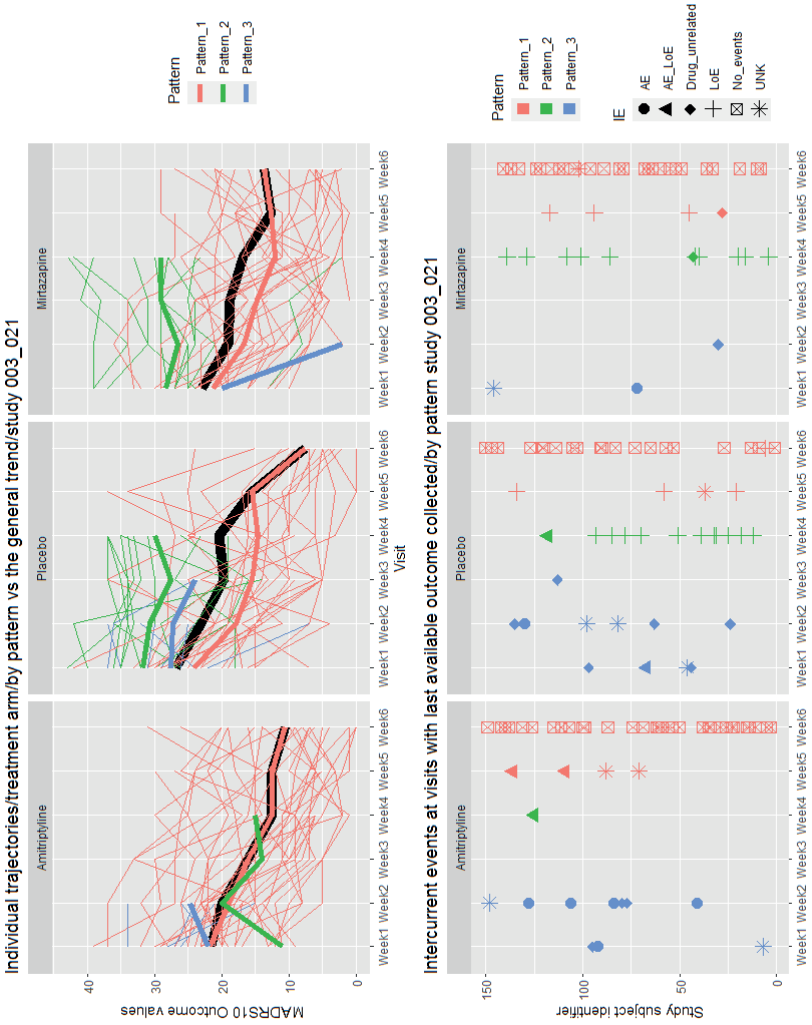

Individual trajectories/treatment arm/by pattern vs the general trend/study 003_020
Amitriptyline
Placebo
Mirtazapine
MADRS10 Outcome values
0
20
40
Week1 Week2 Week3 Week4 Week5 Week6
Visit
Pattern
Pattern_1
Pattern_2
Pattern_3
Intercurrent events at visits with last available outcome collected/by pattern study 003_020
Study subject identifier
0
50
100
Last visit with available outcome
IE
AE
AE_LoE
Drug_unrelated
Efficacy
LoE
No_events
UNK

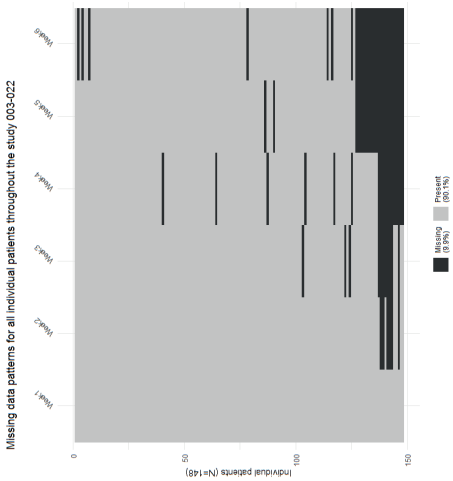

Missing data patterns for all individual patients throughout the study 003-021
Individual patients (N=139)
Missing (24.1%)
Present (75.9%)

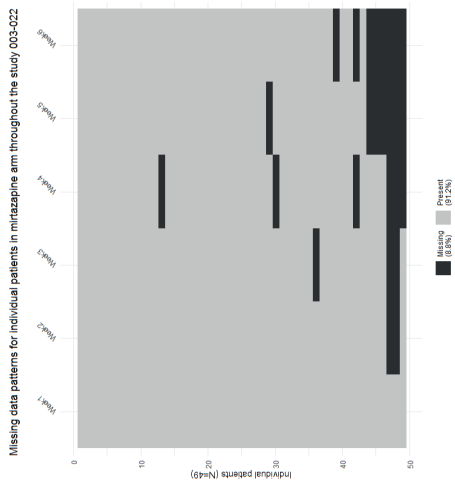

Missing data patterns for individual patients in mirtazapine arm throughout the study 003-021
Individual patients (N=44)
Missing (19.7%)
Present (80.3%)

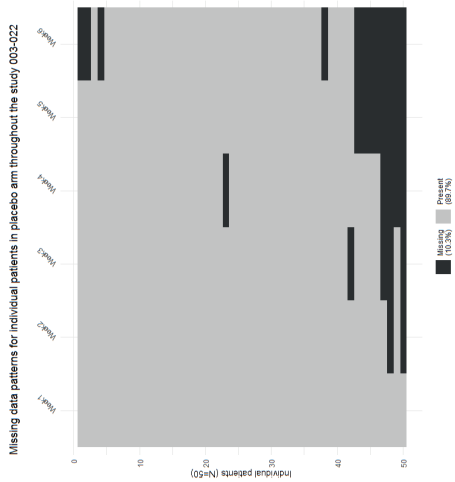

Missing data patterns for individual patients in placebo arm throughout the study 003-021
Individual patients (N=48)
Missing (29.9%)
Present (70.1%)

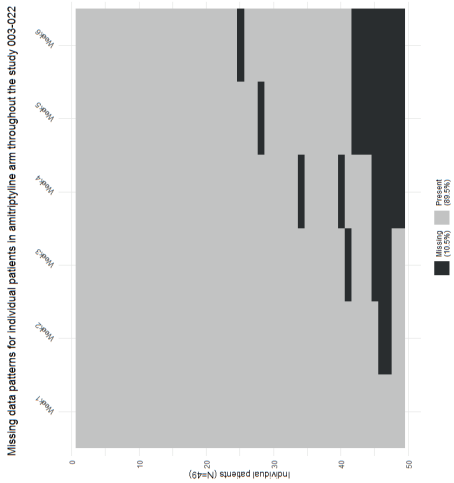

Missing data patterns for individual patients in amitriptyline arm throughout the study 003-021
Individual patients (N=47)
Missing (22.3%)
Present (77.7%)

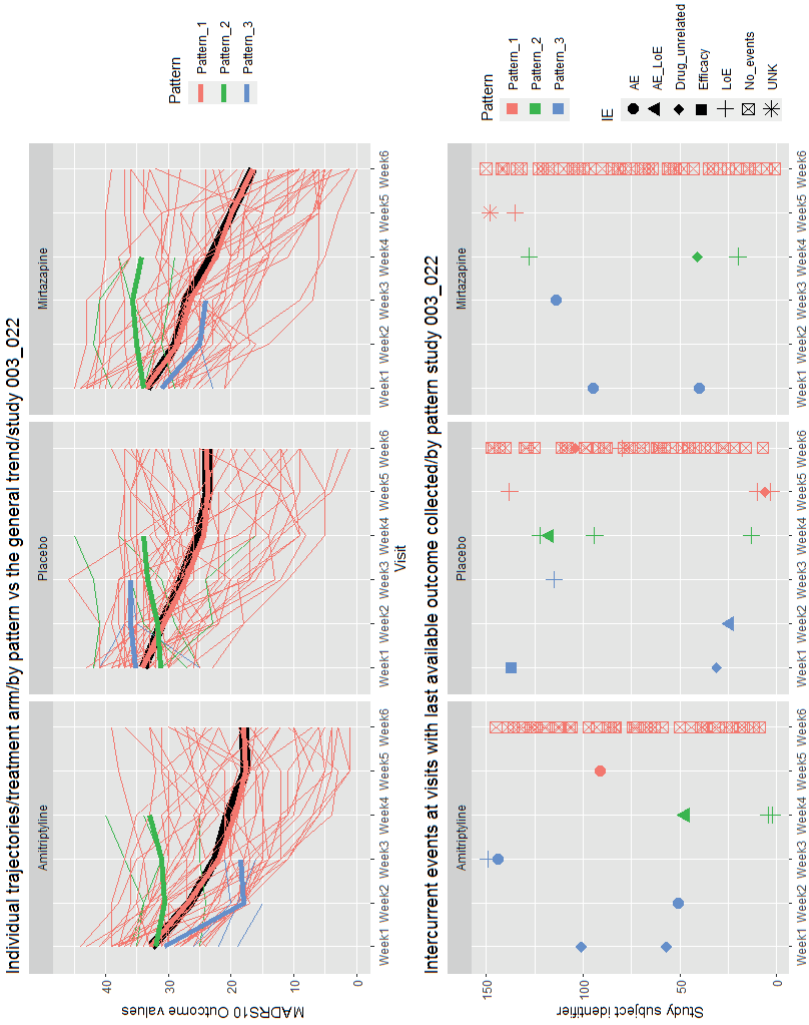

Individual trajectories/treatment arm/by pattern vs the general trend/study 003_021
Amitriptyline
Placebo
Mirtazapine
MADRS10 Outcome values
Visit
Pattern
Pattern_1
Pattern_2
Pattern_3
Intercurrent events at visits with last available outcome collected/by pattern study 003_021
Study subject identifier
Last visit with available outcome
IE
AE
AE_LoE
Drug_unrelated
LoE
No_events
UNK


Missing data patterns for all individual patients throughout the study 003-022
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
0
50
100
150
Individual patients (N=148)
Missing (9.9%)
Present (90.1%)


Missing data patterns for individual patients in mirtazapine arm throughout the study 003-022
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
0
10
20
30
40
50
Individual patients (N=49)
Missing (8.8%)
Present (91.2%)


Missing data patterns for individual patients in placebo arm throughout the study 003-022
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
0
10
20
30
40
50
Individual patients (N=50)
Missing (10.3%)
Present (89.7%)


Missing data patterns for individual patients in amitriptyline arm throughout the study 003-022
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
0
10
20
30
40
50
Individual patients (N=49)
Missing (10.5%)
Present (89.5%)


Individual trajectories/treatment arm/by pattern vs the general trend/study 003_022
Amitriptyline
Placebo
Mirtazapine
MADRS10 Outcome values
Visit
Pattern
Pattern_1
Pattern_2
Pattern_3
Intercurrent events at visits with last available outcome collected/by pattern study 003_022
Study subject identifier
Last visit with available outcome
IE
AE
AE_LoE
Drug_unrelated
Efficacy
LoE
No_events
UNK

Appendix 3 (Supporting information S3): Treatment effect results and corresponding estimand, and by study and analysis method

Study ID	No.	Analysis and imputation method	Number of patients	Point estimate for mean difference between mirtazapine and placebo.*	95% Confidence interval for treatment effect estimate
003-002	Timing of outcome data collection (6 time points) following randomisation: Weeks: 1, 2, 3, 4, 5 and 6				
	1	ANCOVA on complete cases	N1=28 N0=19	-3.43	-10.27, 3.40
	2	ANCOVA following LOCF imputation	N1=44 N0=44	-6.85	-11.39, -2.29
	3	MMRM following LOCF imputation	N1=44 N0=44	-6.47	-10.05, -2.89
	4	ANCOVA following multiple imputation	N1=44 N0=44	-4.41	-9.79, 0.97
	5	MMRM without imputation	N1=44 N0=44	-4.48	- 8.71, -0.24
	6	MMRM following J2R imputation	N1=44 N0=44	-3.68	-7.70, 0.34
	7	PMMM: patterns based on drop out times and type of missing data	N1=44 N0=44	-3.71	-8.11, 0.69
	8a	Propensity score for LoE discontinuations	N1=# N0=#	#	#
	8b	Propensity score for AE discontinuations	N1=# N0=#	#	#
	8c	Propensity score for ALL discontinuations	N1=39 N0=35	-4.72	-9.28, -0.16
84023	Timing of outcome data collection (3 time points) following randomisation: Weeks: 2, 4 and 6				
	1	ANCOVA on complete cases	N1=34 N0=29	-1.93	-7.61, 3.75
	2	ANCOVA following LOCF imputation	N1=54 N0=51	-3.99	-8.80, 0.83
	3	MMRM following LOCF imputation	N1=54 N0=51	-3.92	-7.96, 0.12
	4	ANCOVA following multiple imputation	N1=54 N0=51	-3.14	-8.51, 2.24
	5	MMRM without imputation	N1=54 N0=51	-2.95	-7.19, 1.29
	6	MMRM following J2R imputation	N1=54 N0=51	-2.49	-6.30, 1.32
	7	PMMM: patterns based on drop out times and type of missing data	N1=54 N0=51	-4.12	-8.30, 0.06
	8a	Propensity score for LoE discontinuations	N1=# N0=#	#	#

	8b	Propensity score for AE discontinuations	N1=# N0=#	#	#
	8c	Propensity score for ALL discontinuations	N1=53 N0=50	-3.21	-7.51, 1.08
85027	**Timing of outcome data collection (3 time points) following randomisation: Weeks: 2, 4 and 5**				
	1	ANCOVA on complete cases	N1=48 N0=49	-3.86	-7.70, -0.03
	2	ANCOVA following LOCF imputation	N1=63 N0=61	-3.94	-7.43, -0.45
	3	MMRM following LOCF imputation	N1=63 N0=61	-3.86	-7.20, -0.53
	4	ANCOVA following multiple imputation	N1=63 N0=61	-3.95	-7.52, -0.39
	5	MMRM without imputation	N1=63 N0=61	-4.09	-7.60, -0.59
	6	MMRM following J2R imputation	N1=63 N0=61	-3.32	-6.58, -0.06
	7	PMMM: patterns based on drop out times and type of missing data	N1=63 N0=61	-4.00	-7.46, -0.55
	8a	Propensity score for LoE discontinuations	N1=# N0=#	#	#
	8b	Propensity score for AE discontinuations	N1=# N0=#	#	#
	8c	Propensity score for ALL discontinuations	N1=60 N0=51	-2.79	-6.36, 0.78
003-020	**Timing of outcome data collection (6 time points) following randomisation: Weeks: 1, 2, 3, 4, 5 and 6**				
	1	ANCOVA on complete cases	N1=25 N7=26 N0=25	-6.33 -7.87	-11.58, -1.08 -13.03, -2.71
	2	ANCOVA following LOCF imputation	N1=39 N7=38 N0=37	-6.13 -7.32	-10.38, -1.88 -11.53, -3.10
	3	MMRM following LOCF imputation	N1=39 N7=38 N0=37	-5.93 -7.23	-9.39, -2.47 -10.68, -3.78
	4	ANCOVA following multiple imputation	N1=39 N7=38 N0=37	-6.94 -8.07	-11.49, -2.39 -12.33, -3.80
	5	MMRM without imputation	N1=39 N7=38 N0=37	-5.98 -7.84	-9.79, -2.18 -12.35, -3.33
	6	MMRM following J2R imputation	N1=39 N7=38 N0=37	-1.85 -6.28	-5.35, 1.67 -9.79, -2.77
	7	PMMM: patterns based on drop out times and type of missing data	N1=39 N7=38 N0=37	-6.92 -9.03	-10.92, -2.94 -13.03, -5.03
	8a	Propensity score for LoE discontinuations	N1=# N7=# N0=#	# #	# #

	8b	Propensity score for AE discontinuations	N1=# N7=# N0=#	# #	# #
	8c	Propensity score for ALL discontinuations	N1=30 N7=30 N0=30	-6.17 -7.98	-10.65, -1.69 -12.39, -3.58
003-021	**Timing of outcome data collection (6 time points) following randomisation: Weeks: 1, 2, 3, 4, 5 and 6**				
	1	ANCOVA on complete cases	N1=26 N7=32 N0=21	6.28 2.61	2.12, 10.44 -1.40, 6.62
	2	ANCOVA following LOCF imputation	N1=44 N7=47 N0=48	-0.90 -5.43	-5.38, 3.58 -9.83, -1.04
	3	MMRM following LOCF imputation	N1=44 N7=47 N0=48	-0.78 -5.36	-4.47, 2.91 -8.99, -1.74
	4	ANCOVA following multiple imputation	N1=44 N7=47 N0=48	1.91 -1.58	-1.85, 5.66 -5.50, 2.33
	5	MMRM without imputation	N1=44 N7=47 N0=48	4.60 -1.69	0.41, 8.79 -5.75, 2.37
	6	MMRM following J2R imputation	N1=44 N7=47 N0=48	-2.97 -1.50	-6.74, 0.81 -5.23, 2.22
	7	PMMM: patterns based on drop out times and type of missing data	N1=44 N7=47 N0=48	2.96 -2.57	-1.19, 7.10 -6.89, 1.76
	8a	Propensity score for LoE discontinuations	N1=# N7=# N0=#	# #	# #
	8b	Propensity score for AE discontinuations	N1=# N7=# N0=#	# #	# #
	8c	Propensity score for ALL discontinuations	N1=9 N7=14 N0=10	4.22 2.58	-3.07, 11.52 -4.35, 9.50
003-022	**Timing of outcome data collection (6 time points) following randomisation: Weeks: 1, 2, 3, 4, 5 and 6**				
	1	ANCOVA on complete cases	N1=41 N7=40 N0=38	-7.81 -5.97	-12.79, -2.82 -10.94, -0.99
	2	ANCOVA following LOCF imputation	N1=49 N7=49 N0=50	-7.57 -6.11	-12.20, -2.93 -10.74, -1.49
	3	MMRM following LOCF imputation	N1=49 N7=49 N0=50	-7.89 -6.10	-11.37, -4.42 -9.57, -2.64
	4	ANCOVA following multiple imputation	N7=49 N1=49 N0=50	-6.17 -5.66	-10.85, -1.49 -10.44, -0.89

	5	MMRM without imputation	N1=49 N7=49 N0=50	-8.09 -6.26	-11.49, -4,69 -9.68, -2.83
	6	MMRM following J2R imputation	N1=49 N7=49 N0=50	-4.96 -5.83	-8.29, -1.63 -9.16, -2.50
	7	PMMM: patterns based on drop out times and type of missing data	N1=49 N7=49 N0=50	-8.16 -5.97	-11.74, -4.58 -9.54, -2.40
	8a	Propensity score for LoE discontinuations	N1=# N7=# N0=#	# #	# #
	8b	Propensity score for AE discontinuations	N1=# N7=# N0=#	# #	# #
	8c	Propensity score for ALL discontinuations	N1=34 N7=40 N0=34	-6.17 -4.61	-10.61, -1.73 -8.82, -0.41

#The stratum of interest could not be identified based on available measured covariates
**For studies 003-020, 003-021 and 003-022 the treatment differences are between mirtazapine and placebo (above), and between amitriptyline and placebo (below)*
[a]end of trial is at six weeks for studies 003-002, 84023 and 003-020, 003-021 and 003-022, and at five weeks for study 85027
N7= amitriptyline, N1= mirtazapine, N0= placebo

CHAPTER 5

Data-generating models of longitudinal continuous outcomes and intercurrent events to evaluate estimands

Mitroiu M, Teerenstra S, Oude Rengerink K, Pétavy F, Roes KCB. Data-generating models of longitudinal continuous outcomes and intercurrent events to evaluate estimands

Under review

Abstract

Introduction/Background: We aimed to develop and evaluate data-generating models to jointly simulate outcomes and intercurrent events for randomised clinical trials to enable assessment of properties of estimands.

Methods: We propose four data-generating models for the joint distribution of longitudinal continuous clinical outcomes and intercurrent events under the scenario where they are observable: a selection model, a pattern-mixture mixed model, a shared-parameter model and a joint model of longitudinally observed clinical outcomes and a survival model for intercurrent events. We present a case study in a short-term depression trial with repeated measurements of continuous outcomes and two types of intercurrent events, and compare the four proposed data-generating models.

Results: All four data-generating models can simulate different types of intercurrent events, their timing, and their associated longitudinal outcomes. These can be used to match envisaged patterns of intercurrent events and outcomes informed by prior available clinical trial data. For a given intercurrent event, the Shared-Parameter and Joint model could associate only similar longitudinal profiles, while the Selection Model and Pattern-Mixture model could allow more variation in associated profiles

Conclusion: The four proposed data-generating models can simulate randomised clinical trial data with associated clinical outcomes and intercurrent events. They can be used for simulations to evaluate different estimands and to investigate their properties in-depth. The pattern-mixture mixed model and selection model are more flexible than the shared-parameter models. Understanding and simulating the association between outcomes and intercurrent events may be a solid steppingstone towards optimal implementation of the estimands framework for clinical trials design, analysis, and interpretation.

Introduction

The ICH E9(R1) estimands addendum became public in December 2019, started to be adopted since and is in the process of implementation for regulatory purposes in drug development and evaluation [1–3]. It provides a structured methodological framework for the planning, conduct and interpretation of randomised clinical trials for regulatory evaluation and approval. The main aim is to add clarity and a common understanding between all healthcare stakeholders, of the treatment effects targeted in clinical trials, using estimands.

The estimand is defined as: *"A precise description of the treatment effect reflecting the clinical question posed by the trial objective"*. The estimand is composed of five attributes: treatment, population, variable, population-level summary and strategies to handle intercurrent events [4].

The intercurrent events are defined as: *"Events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest."* These intercurrent events are different from missing data, which is defined as: *"Data that would be meaningful for the analysis of a given estimand but were not collected. They should be distinguished from data that do not exist or data that are not considered meaningful because of an intercurrent event."*.

The proposed estimands framework is still in the early stages of regulatory adoption and implementation in clinical trial conduct [2], but triggered ample discussions [5-8]. Healthcare stakeholders saw its potential and started to follow and apply this guidance for clinical trials to be used for regulatory purposes [5–8]. However, the concrete advantages and tangible added value of estimands framework are not established yet. There is still unclarity regarding, for example, which estimands can or should be used in a randomised clinical trial conducted to confirm a certain claim for a therapeutic indication to be approved [9–12]. There is also uncertainty around the comparison of proposed and formulated estimands, whether they can be estimated and how to estimate them. These key issues could be evaluated by means of simulation studies [13]. To be able to perform a proper simulation study to evaluate estimands and their estimation methods in clinical trials, we need data-generating models (DGMs) to simulate outcome data and intercurrent events under clinically plausible joint mechanisms of occurrence.

There are many DGMs available that can be easily implemented if only the clinical outcome has to be modeled, with or without and/or assumptions on the missing data mechanism [14,15]. However, one of the key innovations with estimands is the use of an

association between occurrence of intercurrent events and clinical outcome. This association needs to appropriately enter the DGM, to allow an enhanced understanding of estimands and adequate simulation. In this research, we focus on DGMs that can model this association between outcomes and intercurrent events.

To our knowledge, there are no published articles that systematically investigate how to model the association between outcomes and intercurrent events. A search query on 9 November 2021 in PubMed of the terms {"generate"/ "simulate" AND "outcomes" AND "intercurrent events" / "post-randomis/zation events"}, and screening of the very few articles suggested, resulted in zero articles that undertook this objective. We restrict this research to DGMs for randomised clinical trials in which continuous outcomes are planned to be collected longitudinally over time (repeated measurements at protocolled visits. We showcase the four proposed DGMs with a short-term anti-depression trial.

Research questions:

1. What data-generating models can be used to jointly generate outcomes and intercurrent events for a generic phase III trial in simulation studies, while modeling the association between them?
2. What are the differences between these data-generating models regarding the model specification and performance in simulating the target trial?

Methods

We present a case study in a short-term depression trial with repeated measurements of continuous outcomes and two types of intercurrent events, and propose four data-generating models to jointly generate outcomes and intercurrent events.

Box 1: Case study in short-term major depressive disorder trials

We consider a two-arm, randomised placebo-controlled, single center, trial of six weeks duration with superiority objective. To achieve a realistic case study where we know the full data, i.e., the intercurrent events and all the outcomes before and after the intercurrent events, we fitted a longitudinal generalized least squares model to an existing dataset from a depression trial (study 003-002), we used this model to inform the final postulated model (Table 2). Based on the same depression trial, we formulated the rules for intercurrent events.

The clinical outcome is MADRS10 and is planned to be measured weekly for six weeks. We simulated 500 trials [16] to allow for approximately 0.1 precision (+/-1.96*SE) in the estimate of treatment effect from the simulated trials (please see supplemental material). Assumed treatment effect (δ) at end of trial at six weeks was -3.5 points on MADRS10 scale, a standard deviation (SD) of outcomes at week 6 of 7.5, to achieve 90% power ($1 - \beta$) and maximum accepted type 1 error (α) of 0.025 (for a one-sided t-test) resulted in needing 95 patients per arm.

We assume early separation in the outcome between the arms with the treatment effect maintained to the end of trial. Intercurrent events of interest are treatment discontinuation due to LoE at week 3 and due to AE at week 2. These assumptions are based on study 003-002 (source trial for the case study). The targeted proportion of LoE at trial level is 35% (21% in control, 14% in experimental), the targeted proportion of AE at trial level is 15% (5% in control, 10% in experimental). The patients not experiencing any intercurrent events are randomly allocated evenly at trial level (they are the remainder of simulated subjects that do not experience any intercurrent event as they do not fit in the rules). A downward time trend of outcomes is expected (improvement), based on the source trials used to extract the DGMs parameters.

The main steps for constructing the DGM are: 1) Investigate available data from completed trials (source trials), 2) postulate models corresponding to relevant clinical assumptions and extract model parameters, 3) Simulate trials with outcomes and intercurrent events, 4) Verify the generated datasets for concordance with the target trial.

In our simulations we followed recommendations for designing and conducting simulation studies used to evaluate statistical methods for medical research [16,17].

The verification of longitudinal outcomes and intercurrent events generated with each of the four DGMs, consists of three steps: verification of the longitudinal outcomes (compare numeric values of model parameters against predefined assumed true parameters and visually inspect the trajectories), of the intercurrent events (compare the type, timing and percentages tabulated against predefined type, timing and percentages). Each verification is performed against the DGM-specific procedure and parameters. We compared the four proposed data-generating models.

All simulations were conducted in RStudio (Version 1.4.1717, "Juliet Rose" (df86b69e, 2021-05-24) for macOS) with relevant packages [18–26]. The verbose, annotated R script developed and used for the simulations for all four DGMs was double-checked by a researcher independent of this project. For ease of application of the proposed DGMs, we provide this R script to facilitate simulations of envisaged trials [27].

The simulation would typically start from available data for outcomes and intercurrent events from an already conducted trials (source trials). Models underlying the DGMs can be fitted on these data to estimate DGM's parameters and used to simulate outcomes and intercurrent events in the target trial.

Four models (DGMs) to jointly generate outcomes and intercurrent events

We focus on a randomised, parallel-group, placebo-controlled, two-armed trial with repeated measurements of continuous outcomes, and with several intercurrent events.

The clinical setting involves Y_{ij} repeated measurements of continuous outcomes, the vector $Y_i = \{Y_{ij}\}$, $i = 1, .., n$ patients, j = visit number, with X_i associated baseline covariates.

Let IE_{ij} be the specific intercurrent event indicator for Y_{ij}, with value 1 if the intercurrent event is recorded at visit j to have been experienced between visit $j - 1$ and j in the time interval $(j - 1, j]$, and with value 0 if the intercurrent event is not experienced in the interval $(j - 1, j]$, denoted by vector $IE_i = \{IE_{ij}\}$. This means that the intercurrent event must have occurred (it is an event) before or at the latest, simultaneously when Y_{ij} was measured at visit j. This is to be closely evaluated because the intercurrent event IE_{ij} must have affected the outcome Y_{ij} (interpretation or existence, as per definition in E9(R1)), otherwise the intercurrent event IE_{ij} is not an intercurrent event for Y_{ij}, but possibly for Y_{ij+1}.

The vector Y_i^b denotes the set of outcome values for patient i before the intercurrent event IE_{ij} is experienced, and Y_i^a the set of outcome values for patient i after the intercurrent event IE_{ij} is experienced. $\{Y_{ij}\} = \{Y_i^b, Y_i^a\}$. If $IE_{ij} = 0$, then Y_{ij} is not split into Y_{ij}^{ab} and Y_{ij}^a.

Full data $Z_{ij} = \{Y_{ij}, IE_{ij}\}$, outcomes and intercurrent events, are generated based on a DGM for the simulation study.

In all four proposed DGMs, there is a need of for a model for the joint distribution of Y_{ij} and IE_{ij}, with density $f(IE, Y \mid X, \theta)$, indexed by parameters $\theta = (\gamma, \phi)$, where $gamma$ (γ)

characterizes the model for the outcomes, and *phi* (ϕ) characterizes the model for the intercurrent events.

We specify the joint likelihood of outcomes and intercurrent events using similar approaches as are used for missing data modeling.

High-level descriptions of the DGMs

The Selection Model (SM) conditions the occurrence of the intercurrent event on the values of the longitudinal outcomes. The Pattern-Mixture Mixed Model (PMMM) conditions the longitudinal outcomes on the intercurrent event (type, timing, etc.). The shared-parameter model (SPM) and joint model of longitudinal outcomes and survival model for intercurrent events (hereafter referred to as "JM"), do not specifically condition the outcomes directly on the intercurrent events or the intercurrent events on the outcome, but both are conditional on shared random effects.

1. Selection model (SM)

The principle of the selection model is to first generate the longitudinal outcome data (Y_{ij}) and then 'select' based on a rule which subjects experience the intercurrent events based on these outcomes, treatment, and type of intercurrent event.

Outcomes Data Generating Model for Y_{ij} via marginal model

From longitudinal outcomes of other phase II or phase III studies, the trajectories parameters and variance-covariance matrix of outcomes to generate correlated residuals ε_{ij} (errors) can be informed. These data can be used to model the means (β_{1j} and/or $\beta_{2jTreat}$) at different visits *j* for the longitudinal outcome trajectories (Table 1).

Table 1. Outcomes DGM specification and notation (Selection Model)

$Y_{ij} = (\beta_{1j} + \varepsilon_{ij}) + \beta_{2jTreat} x\ Treatment_i$
Y_{ij} = repeated measurements of continuous outcome for patient *i* at baseline (*j*=0) or follow-up visit *j* ($e.g., \ j = 1, \ldots, 6$) $\varepsilon_{ij} \sim N\left(0, \Sigma^2_{\varepsilon_{ij}}\right)$ $Treat_i$ = randomised treatment for patient *i*, indicator is 0 for control and 1 for treatment ε_{ij} = correlated residuals for patient *i* at baseline (0) and follow-up visit *j (week 1 to week 6)* $\Sigma^2_{\varepsilon_{ij}}$ = unstructured covariance matrix of residuals ε_{ij}

Other models can be used to generate the longitudinal outcome data, e.g. a model with random effects model (e.g. random intercept and/or random slope) if this describes the observed data well.

Intercurrent events generating model (IEGM for IE_{ij})

In the selection model approach, IE_{ij} are only dependent on the longitudinal outcomes Y_{ij}.

The selection model specifies the joint distribution of Y_i and IE_i through models for the distribution of Y_i and the conditional distribution of IE_i given Y_i.

$$f(IE_i, Y_i|X_i,\gamma\,,\phi) = f_Y(Y_i|X_i,\gamma\,)f_{IE|Y}(IE_i|X_i, Y_i, \phi)\,,$$

where $\theta = (\gamma, \phi)$, set of parameters modeling the intercurrent events and outcomes.

The underlying assumption for this DGM is that intercurrent events and their occurrence are directly connected with the clinical outcomes – either observed before or after the occurrence of the intercurrent event. Based on clinical input, for instance, if a patient does not improve by a certain difference at end of trial, then lack of efficacy is experienced at the visit mid-trial. For adverse events (AE), it could be differently, if a patient does improve but insufficiently from start to mid-trial to compensate for the AE, then the patient stops treatment for AE recorded at mid-trial.

2. Pattern-mixture mixed model (PMMM)

The pattern mixture model follows a different factorization of the joint model for IE_{ij} and Y_{ij}. The longitudinal outcomes and the intercurrent events follow a joint distribution, through the marginal distribution of IE_i and the conditional distribution of Y_i given IE_i:

Generate the full data $Z_{ij} = \{Y_{ij}, IE_{ij}\}$ as follows:

$$f(IE_i, Y_i|X_i, \nu, \delta) = f_{IE}(IE_i|X_i, \delta)f_{Y|IE}(Y_i|X_i, IE_i, \nu)\,,$$

$$(Y_{ij}\,|IE_{ij} = k) \sim N(\mu(k), \Sigma(k)).$$

Namely, the profiles associated with a type and timing of IE together form a pattern (in the sense of pattern mixture). The profiles in this pattern are multivariate normal distributed with a mean and covariance matrix specific to each pattern k.

The number of patients in each pattern can be pre-determined (specific fixed proportion of each pattern), stochastic (varying proportions of each pattern) or a mix (varying proportion of each pattern and a specific fixed proportion of a certain or more patterns).

3. Shared-parameter model (SPM)

The repeated measurements (Y_{ij}) of continuous outcomes are modeled using a random effects model, such as a random intercept and/or random slope model (subject-specific model). Then, the intercurrent events follow a logistic regression model (the outcome is an intercurrent event at a particular timepoint *j*) with fixed and random effects. At least some of the random effects in this logistic model and some of the random effects model for the longitudinal outcomes are common, i.e., they are shared. The logistic model gives the probability for patient i to experience the intercurrent event IE at timepoint j conditional on the random effects shared with the longitudinal outcomes model, for example, a linear mixed-effects model. Each subject has latent traits expressed by random effects. Both the longitudinal outcome of the subject and the probability to experience an intercurrent of specific type and timing depend on this latent trait (e.g. frailty).

Outcomes sub-model

The process for repeated measurements of continuous outcomes is as follows:

$$Y_{ij} = X_i^Y \beta^Y + Z_i^Y b_i + \varepsilon_{ij},$$

$$b_i \sim N\left(0, \sigma_b^2\right), \;\; \varepsilon_{ij} \sim N\left(0, \sigma_{\varepsilon_{ij}}^2\right).$$

Intercurrent events sub-model

The intercurrent events occurrence mechanism is "at random". The intercurrent events to be experienced are, conditional on random effects b_i, distributed as

$$logit\left(P\left(IE_{ij}\right)\right) = log\left(\frac{P\left(IE_{ij}\right)}{1 - P\left(IE_{ij}\right)}\right) = X_i^{IE} \beta^{IE} + Z_i^{IE} b_i + \varepsilon_{ij}^{logit}.$$

Here X_i^x is the design matrix for the fixed effects β^x, which could have the same/different fixed effects as used in the linear-mixed effects model to generate Y_{ij}, or other combination of linear predictors. Z_i^x is the design matrix for the random effects b_i. This step can require finetuning for postulated models or models can be fitted on actual data directly.

For each intercurrent event, at particular timepoints *j*, a logit model can be postulated to model its occurrence based on fixed effects (e.g. covariates) and random effects.

The logit models for the estimated probabilities of occurrence of intercurrent events for specific patients at specific timepoints *j* are directly dictated by the patients' subject-specific random intercepts and/or slopes (shared random effects).

4. Joint modeling of repeated measurements and intercurrent events via JM

Outcomes sub-model

In the same manner as in SPM DGM (Shared-Parameter Model) [28–30], the longitudinal profiles are generated via random effects as:

$$Y_{ij} = X_i^Y \beta^Y + Z_i^Y b_i + \varepsilon_{ij} = m_i(t) + \varepsilon_{ij},$$

$$b_i \sim N(0, \sigma_b^2), \quad \varepsilon_{ij} \sim N\left(0, \sigma_{\varepsilon_{ij}}^2\right).$$

Intercurrent events sub-model

Intercurrent events are modeled with a (parametric) survival model. In our case we use a Weibull distribution to draw time to intercurrent event data for each patient:

$$h_i(t|M_i^*(t)) = h_0(t)exp\{\gamma\ \omega_i + m_i^*(t)\},$$

$$S_i(t|b_i) = exp\left(-\int_0^t h_0(s)exp\{\gamma\ \omega_i + \alpha m_i^*(t)\}ds\right).$$

where:

w_i = baseline covariates

γ = parameters for baseline covariates

$Y_i(t) = Y_{ij}$ = longitudinal

$M_i^*(t) = \left\{ m_i^*(s),\ 0 \le s < t \right\}$

$\gamma w_i + \alpha\ m_i^*(t)$ = linear predictor to generate time to intercurrent events

m_i^* = the estimated true longitudinal measurement that contains at least some or all random effects b_i that are also part of the linear predictor for the outcomes Y_{ij};

α is a regularization factor that quantifies the strength of the association between the true longitudinal measurement and the risk of an intercurrent event.

The random effects contained in $m_i^*(t)$ influence the hazard of an intercurrent event for each individual patient i (subject-specific hazard). We use time as a continuous variable, if the intercurrent event falls (likely) between pre-planned visits, the next coming visit in time can be considered to have recorded the intercurrent event.

Trial design features and setup for simulations

The following trial design features need to be set: type of outcome, number and timing of visits, number of trial arms, randomisation ratio, assumed trajectories of patients' outcomes (DGM model parameters) and objective of the study, the expected outcome over time without treatment (natural history) or under control treatment and after stopping investigational treatment, and how they relate with the intercurrent events and the mechanism of occurrence, as well as relevant covariates.

While planning the simulation, the following trial design considerations can be followed in order to decide the settings of trial design features and which of the four proposed DGMs is suited to use: type, timing (e.g. at specific timepoints or spread throughout the trial duration, early, late, uniformly or non-uniformly distributed throughout the trial), percentages at trial level, within arm and the ratios between them at trial and arm level, and how they relate with the outcomes, the dependency on outcomes.

Results

For the case study described below, we simulated the targeted short term major depressive disorder trial with outcomes and intercurrent events using all four proposed DGMs to model the association between outcomes and intercurrent events.

We described for each DGM the considerations for implementation in Appendix 1.

Description of the target trial parameters for the four models

1. Selection Model

We display SM DGM with the deterministic rule implementation. The corresponding graphs and table for SM DGM stochastic implementation are in supplemental material (Appendix 4).

Table 2. Assumed true parameters for trajectories in each arm for the target trial

Parameter name	Parameter value
beta.baseline (β_{10})	29.79
beta_week1 (β_{11})	-1
beta_week2 (β_{12})	-1.5
beta_week3 (β_{13})	-2
beta_week4 (β_{14})	-2.5
beta_week5 (β_{15})	-3
beta_week6 (β_{16})	-3.5
beta_v1_treatment ($\beta_{21Treat}$)	-1
beta_v2_treatment ($\beta_{22Treat}$)	-1.5
beta_v3_treatment ($\beta_{23Treat}$)	-2
beta_v4_treatment ($\beta_{24Treat}$)	-2.5
beta_v5_treatment ($\beta_{25Treat}$)	-3
beta_v6_treatment ($\beta_{26Treat}$)	-3.5

Please see Table 1 for the interpretation of these parameters.

For lack of efficacy, we apply the same rule for both arms based on the within patient difference at visit 6 ($D_{i6} = Y_{i0} - Y_{i6}$). If $D_{i6} < 5$ points on MADRS10 (insufficient efficacy), then treatment discontinuation due to lack of efficacy will be assigned at $j = 3$ for that particular patient.

For adverse events we apply a different rule for each arm based on observed efficacy and the assumed relation between very high efficacy and occurrence of AE.

The IE is assumed to occur between day 8 and day 14 inclusively, and recorded at day 14 visit (see described above in the Selection Model DGM). $D_{i1} = Y_{i0} - Y_{i1}$:

For the experimental arm if $D_{i1} > 3$, then treatment discontinuation due to adverse events at $j = 2$ for that particular patient. This corresponds to a steep decrease in MADRS10 observing good (or too much) efficacy, but the patient experiences AE due to toxicity or accumulation additional to the perceived efficacy.

For the control arm if $D_{i1} \leq -2$, then treatment discontinuation due to adverse events at $j = 2$ for that particular patient, as for control arm patients there is likely no relation

between high efficacy and AE. This corresponds possibly to non-decrease in MADRS10 to some extent, if the patient experiences AE and insufficient efficacy, the benefit to AE-detrimental effect ratio is not positive to stay in the trial, AE being the principal reason. To steer the percentages of adverse events, all control patients that meet this deterministic rule have subsequently a probability $P_{AE_{Treat}}$ to experience the AE. $P_{AE_{Treat}}$ is sampled for each simulated trial from a uniform distribution with mean $proportio\,n_{AE}$ in order to add more variability. This range of the uniform distribution is chosen such that the ratios for LoE:AE within arms are within parameters mentioned above.

For the deterministic implementation of the SM DGM, these probabilities have to be set in accordance with the rules for intercurrent events (as possibly observed in real trials) and with the desired percentages of intercurrent events to be achieved. The maximum percentages of each intercurrent event that can be simulated is limited by the rules for intercurrent events (e.g. how many patients in the trial have a decrease from baseline of less than 5 points on MADRS10 scale).

For the stochastic implementation, the percentages depend on the linear predictor terms. Below are model parameters that were fitted on a very large trial where the profiles and intercurrent events were simulated based on the above SM DGM deterministic implementation. Subsequently, these models will follow stochastically the rules for intercurrent events and corresponding percentages.

Table 3. Parameters extracted and used for linear predictors in the logit models for intercurrent events for the SM DGM, stochastic implementation

Logistic regression model	Parameter name	Parameter value
LoE at trial level	$\phi_0^{LoE_{trial}}$	1.077405
	$\phi_1^{LoE_{trial}}$	-0.355617
AE in experimental arm	$\phi_0^{AE_{exp}}$	-2.370965
	$\phi_1^{AE_{exp}}$	0.272188
AE in control arm	$\phi_0^{AE_{ctrl}}$	-2.78108
	$\phi_1^{AE_{ctrl}}$	-0.39944

ϕ_0 = *intercept*, ϕ_1 = *parameter for the outcome difference of interest according to the rules for intercurrent events*

Unless the rules for LoE and AE exclude each other, there may be competing events in the SMs. In that case the earlier event (the AE in this case) takes priority and will be assigned.

2. Pattern-Mixture Mixed Model

The order of operations for the PMMM DGM is to identify the patterns (i.e., the subsets of patients with a given type and timing of intercurrent event) and estimate the longitudinal parameters for each pattern. In concordance with the proportions of each pattern at trial level and in concordance with the overall trial parameters, the weighting of these patterns must be refined, such that the weighted averages correspond when stacking (see weighted average formula below) together the patterns and comprising the (full) target trial.

Table 4. Model parameters for each pattern of the PMMM DGM

Parameters	**Proportions:**	**0.35**	**0.15**	**0.50**	**Check at parameter level**
	Trial level	**LoE in both arms** **0.14 in exp** **0.21 in ctrl**	**AE in both arms** **0.10 in exp** **0.05 in ctrl**	**No IE in both arms** **1:1 random unconstrained**	
Beta0	**29.79**	**29.79**	**29.79**	**29.79**	29.79
Beta1	-1	**1**	**0.1**	**-2.75**	-1.01
Beta2	-1.5	**1.2**	**0.1**	**-3.85**	-1.49
Beta3	-2	**1.2**	**0.1**	**-4.85**	-1.99
Beta4	-2.5	**1.8**	**0.3**	**-6.35**	-2.5
Beta5	-3	**2.8**	**1**	**-8.25**	-2.995
Beta6	-3.5	**0**	**0**	**-7**	-3.5
Beta7	-1	**0.1**	**-2**	**-0.7**	-1.006756703
Beta8	-1.5	**0.8**	**-4**	**-1.15**	-1.508520608
Beta9	-2	**-0.2**	**-4**	**-1.5**	-1.992814326
Beta10	-2.5	**-0.5**	**-3.4**	**-2.25**	-2.504744698
Beta11	-3	**-0.5**	**-5**	**-2**	-2.986739496
Beta12	-3.5	**-1**	**-5**	**-4**	-3.486994798

$$weighted\ average\ for\ treatment\ effect\ at\ week\ j = \frac{\sum_{p=1}^{m} \beta^{j}_{pattern_p^{exp}} \ x\ W_{pattern_p^{exp}}}{\sum_{p=1}^{m} W_{pattern_p^{exp}}} - \frac{\sum_{p=1}^{m} \beta^{j}_{pattern_p^{control}} \ x\ W_{pattern_p^{control}}}{\sum_{p=1}^{m} W_{pattern_p^{control}}}$$

w_p = *weight or proportion of a pattern of intercurrent events in a certain arm*

m = *a certain pattern of intercurrent events or behaviour* ("AE", "LoE" or "No intercurrent events", respectively)

exp = experimental arm

3. Shared-Parameter Model

Below are the parameters estimated, refined and used for the linear mixed-effects and logistics regression models used to simulate trials with the SPM DGM.

Table 5. Model parameters for SPM DGM
Linear mixed-effects model

Parameter name	Parameter value
Fixed effects	
β_0	29.374914
β_1	-0.529752
β_2	-0.577650
Random effects	
$b_i \sim N(0, \sigma_b^2)$	$\sigma_b^2 = \begin{pmatrix} 21.140783 & 1.307053 \\ 1.307053 & 1.044141 \end{pmatrix}$
Random error	
$\varepsilon_i \sim N(0, \sigma_i^2)$	$\sigma = 3.285955$

Table 6. Model parameters for logit models of SPM DGM

Logistic regression model	Parameter name	Parameter value
LoE at trial level	β_0	-0.0877193
	c_1	-0.60629
	α	1
AE in experimental arm	β_0	-1.353383
	α	1
AE in control arm	β_0	-2.155388
	α	1

$logit\left(P\left(LoE_{ij}\right)\right) = \left(\beta_0 + b_{i1}\, x\, \alpha\right) + c_1\, x\, Treat_i$, $logit\left(P\left(AE_{ij}^{Treat}\right)\right) = \beta_0 + b_{i1}\, x\, \alpha$, $logit\left(P\left(AE_{ij}^{Control}\right)\right) = \beta_0 + b_{i1}$, $b_{i1} = random\ slopes$ (Same as in the linear-mixed effects model for the continuous outcomes).

For each patient a random intercept and slope are drawn. Together with the fixed effects these generate the longitudinal profile and the probability to experience an intercurrent event. Thus, the intercurrent event data are generated by drawing from a Bernoulli distribution with mean p = individual probability to experience the intercurrent event IE_{ij} according to Table 6.

$$(IE_{ij} \mid b_{ix}) \sim Bernoulli\left(P\left(IE_{ij}\right)\right)$$

$$(LoE_{i3} \mid b_{i1}) \sim Bernoulli\left(P(LoE_{i3})\right)$$

The treatment discontinuation due to LoE to be experienced by patient i at timepoint $j = 3$ is Bernoulli distributed with mean equal to the probability of patient i to experience LoE at timepoint 3. This probability of each patient i is given by the logit models including the individual random effects (shared with the longitudinal outcomes model).

Depending on the variances of random effects, their distribution and symmetry relative to the position on the sigmoid logit curve, and on their weight in the linear predictor of the logit models for intercurrent events, the obtained percentages of intercurrent events may vary. We conducted checks with small weights in the linear predictor and obtained exactly the percentages we desired. Another finetune to be made is an adjustment in the model such that the percentage of LoE is as desired and due to the competition with AE. Hence, an adjustment can be made, approximately, in order to obtain the targeted percentages of LoE. Precise percentages can be achieved by decreasing variance of random effects, decreasing their weight in the linear predictor or by increasing the number of patients and/or simulated trials.

4. Joint Model

Below are the parameters estimated, refined and used for linear predictors in the survival models for intercurrent events used to simulate trials with the SPM DGM. The linear mixed-effects model parameters are the same as used in the SPM DGM.

Survival model parameters

Table 7. Parameters of the survival models for intercurrent events in the JM DGM

Survival model	Parameter name	Parameter value
LoE at trial level	β_0	2.8789
	c_1	0.3111
	α	−0.0279
	λ	1.7618
	ν	1.7618
AE at trial level	β_0	3.7550
	c_2	−0.5394
	α	0.0076
	λ	0.9719
	ν	0.9719

$$linear\ predictor\ lack\ of\ efficacy\ (LP_{LoE}) = \left(\beta_0 + b_{i0}\, x\, \alpha\right) + c_1\, x\, Treat,$$

$$time\ to\ lack\ of\ efficacy = \left(-\frac{log(U)}{\lambda}\ exp\{LP_{LoE}\}\right)^{1/\nu}$$

$$linear\ predictor\ adverse\ events\ (LP_{AE}) = \left(\beta_0 + b_{i0}\, x\, \alpha\right) + c_2\, x\, Treat,$$

$$time\ to\ adverse\ events = \left(-\frac{log(U)}{\lambda}\ exp\{LP_{AE}\}\right)^{1/\nu}$$

To inform the model specification, joint models can be fitted on the source trial data (e.g. Weibull model for survival data) [31]. These models can be then used directly to simulate the intercurrent events data. Another implementation for the survival models would be an iterative process to finetune the parameters of the model (scale, shape) such that the distribution of intercurrent event times is the one envisaged. In particular, patients with no intercurrent event have their event time beyond the trial duration.

Intercurrent events are here generated (drawn from the Weibull distribution), such that most of lack of efficacy is experienced at weeks two and three, the rest of the proportions distributed through the entire duration of the trial. Percentages of intercurrent events are steered by the percentage of patients with their event time beyond the trial duration.

All model parameters for each pattern (each intercurrent event) in the DGMs have been informed by models fitted on actual trial data.

Inspection of outcomes and intercurrent events for the four DGMs

For the visual inspection of outcomes and intercurrent events, the last simulated trial (from the set of $n = 500$ simulated with each DGM) was selected. The clinical outcomes were graphed and inspected to check the longitudinal trajectories for each pattern (lack of efficacy, adverse event, no intercurrent events). A spaghetti plot is provided for a real depression trial [32], that visualizes the longitudinal outcomes (MADRS10) of patients. This is used as an anchor to real-life trial data (not simulated), Figure 1. These simulated trials are displayed by pattern and at trial level for the patterns generated by each DGM (Figure 2 lack of efficacy, Figure 3 adverse event, Figure 4 no intercurrent event), Figure 5 all patterns stacked together).

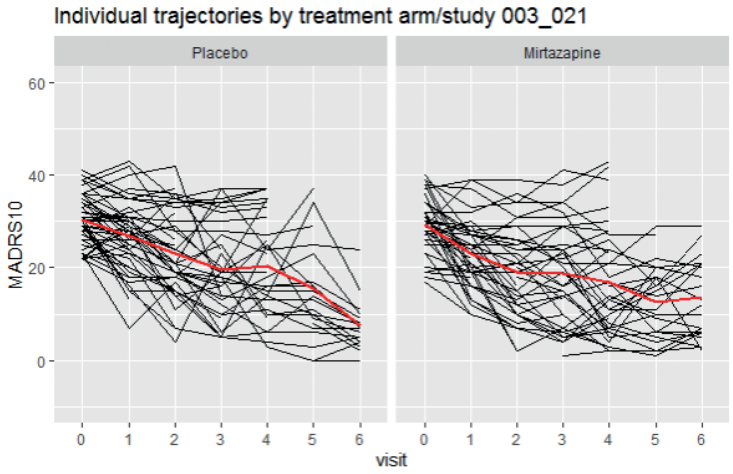


Figure 1. Individual trajectories by treatment arm/study 003-021

Graphs to visualize each pattern in the trials simulated with each DGM

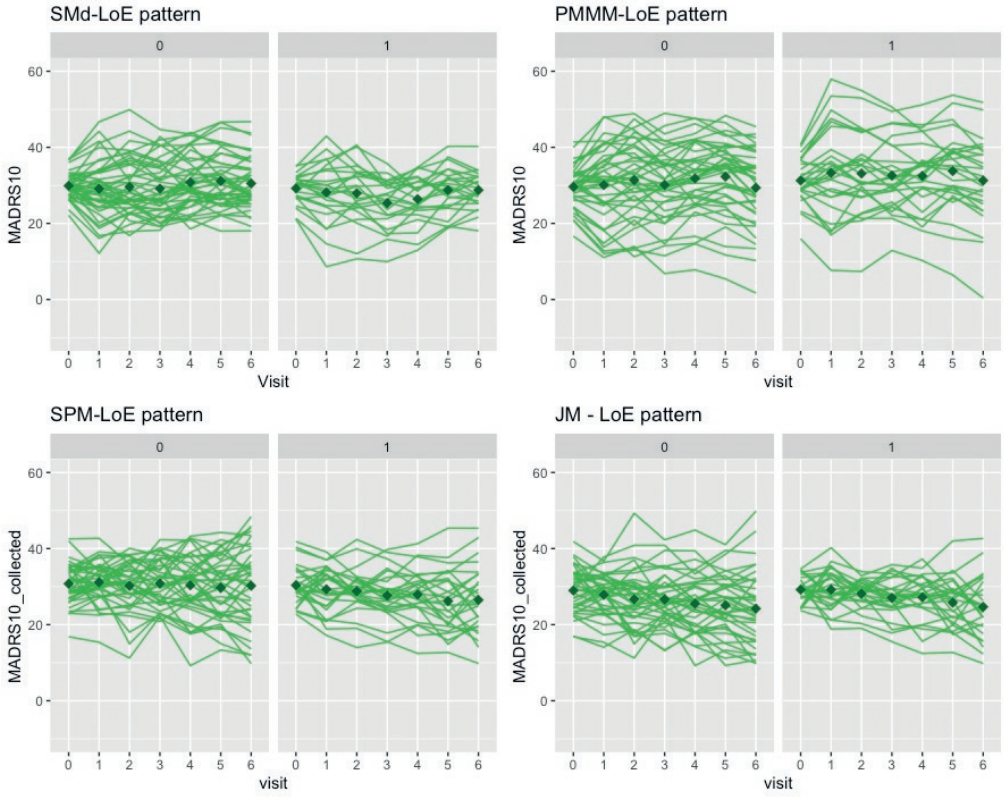


Figure 2. Lack of efficacy patterns by each DGM and arm

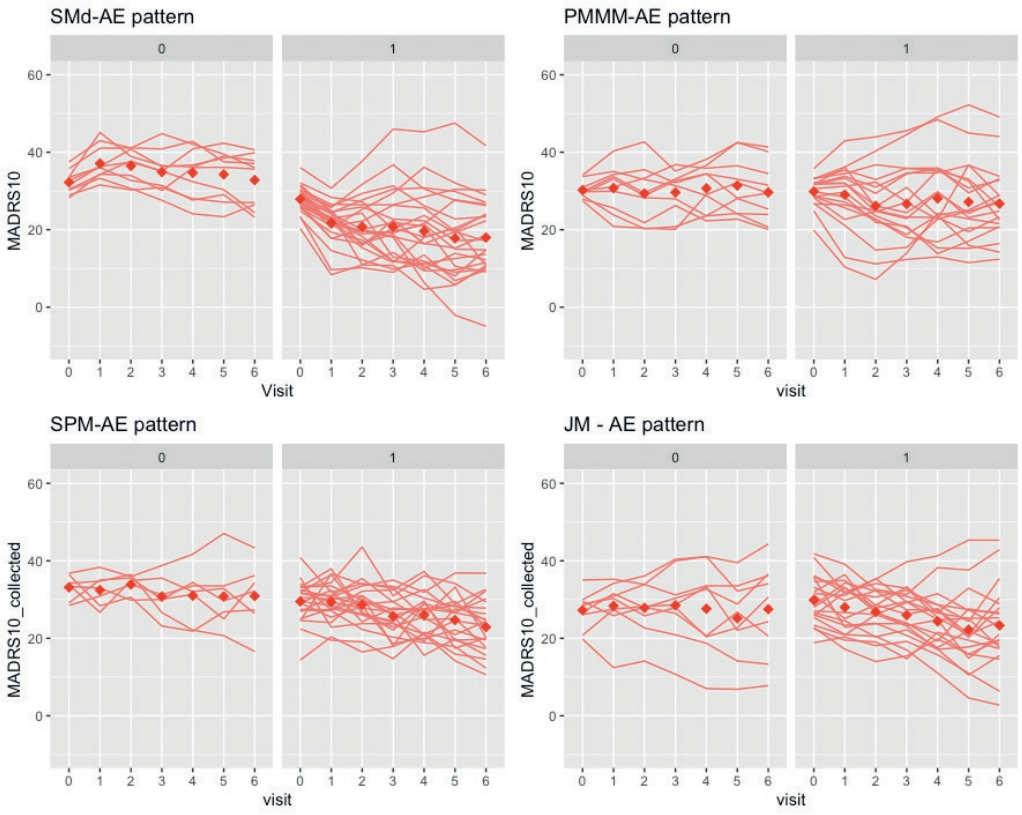


Figure 3. Adverse events patterns by each DGM and arm

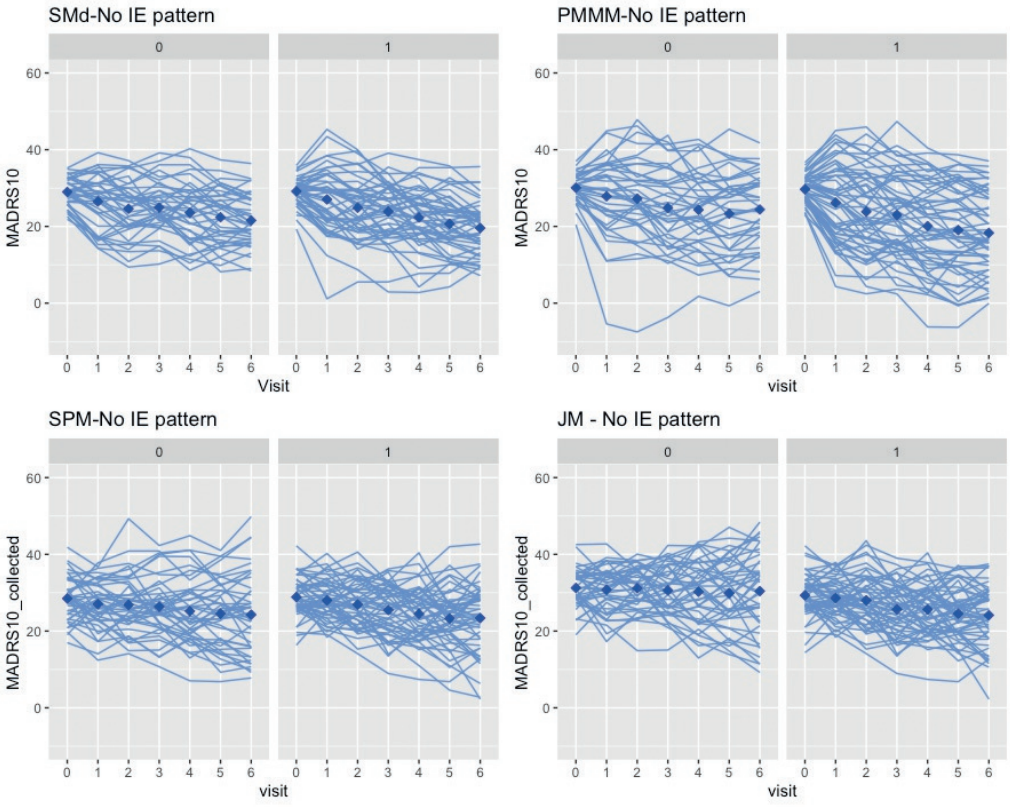


Figure 4. No intercurrent events patterns by each DGM and arm

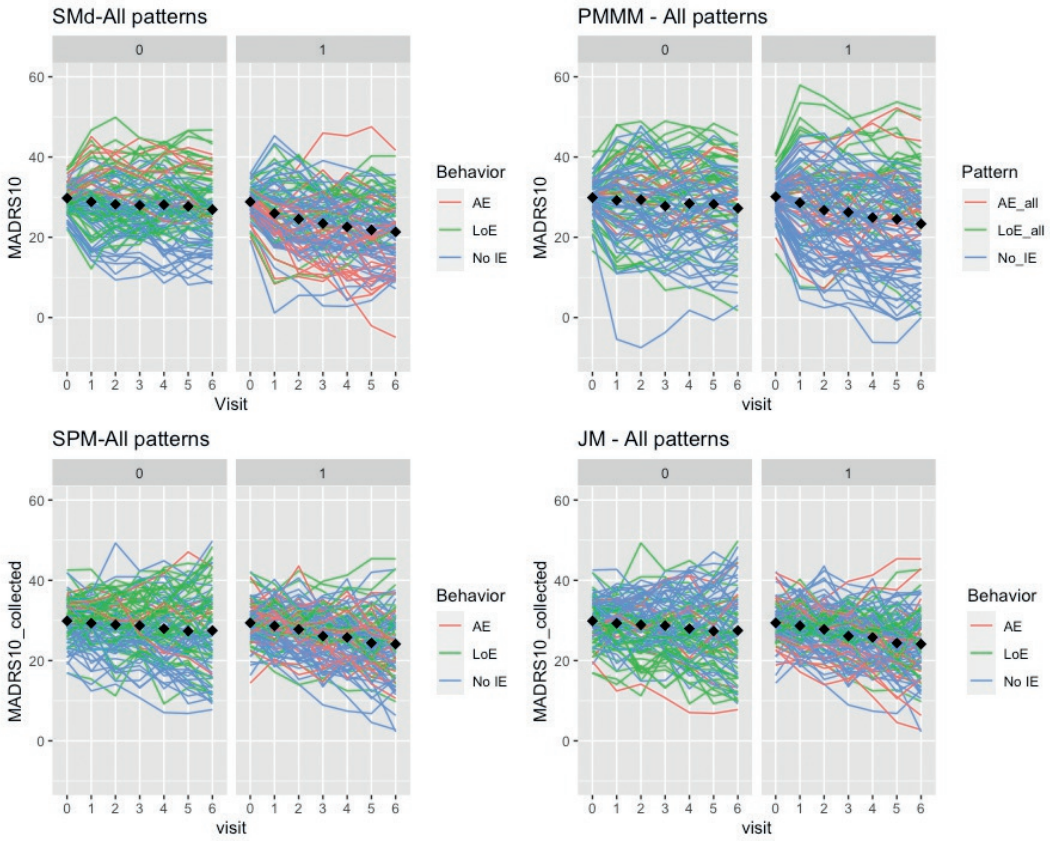


Figure 5. All patterns stacked together as a trial simulated by each DGM and arm

0 = control (placebo), 1 = treatment (mirtazapine)

For the verification of intercurrent events percentages and prespecified ratios within the trial and between arms, we summarise the percentages of intercurrent events by type and timing (Tables 8 a-d).

Table 8a. Descriptive statistics intercurrent events

Selection model DGM - deterministic rule

	Control		Treatment		Total	
Intercurrent event	N	%	N	%	N	%
AE	10.1	5.3	18.1	9.5	28.2	14.9
LoE	39.5	20.8	26.8	14.1	66.3	34.9
Total	49.6	26.1	44.9	23.6	94.5	49.8

Averaged over 500 simulated trials. Trial sample size = 190

Table 8b. Descriptive statistics intercurrent events

Pattern-mixture mixed model DGM

	Control		Treatment		Total	
Intercurrent event	N	%	N	%	N	%
AE	10	5.3	19	10.0	29	15.3
LoE	40	21.1	27	14.2	67	35.3
Total	50	26.4	46	24.2	96	50.6

Averaged over 500 simulated trials. Trial sample size = 190

Table 8c. Descriptive statistics intercurrent events

Shared parameter model DGM

	Control		Treatment		Total	
Intercurrent event	N	%	N	%	N	%
AE	9.9	5.2	19.7	10.4	29.6	15.6
LoE	40.8	21.5	25.6	13.5	66.4	34.9
Total	50.7	26.7	45.3	23.9	96.0	50.5

Averaged over 500 simulated trials. Trial sample size = 190

Table 8d. Descriptive statistics intercurrent events

Joint Model DGM

	Control		Treatment		Total	
Intercurrent event	N	%	N	%	N	%
AE	10.0	5.2	19.3	10.1	29.2	15.4
LoE	41.7	21.9	24.6	12.9	66.2	34.9
Total	51.7	27.1	43.9	23.0	95.4	50.3

Averaged over 500 simulated trials. Trial sample size = 190

Tables 8 a-d. Percentages of intercurrent events for each DGM (a = SM, b = PMMM, c = SPM and d = JM)

Comparison of the DGMs

All four DGMs that model the association of the longitudinal outcomes and intercurrent events can be informed by already conducted trials.

The only DGM that can ensure exact numbers and constant percentages of intercurrent events across all simulated trials, is PMMM. The other three DGMs can ensure approximately the desired percentages of intercurrent events, but still need subtle finetuning in order to do so. Each DGM could be used to simulate multiple intercurrent events, possibly at different timings, for an individual patient.

The longitudinal profiles of the SM and PMMM DGMs are more flexible than SPM and JM. The SPM and JM put restrictions due to the specification and covariance of random effects, and the change in group means. This is not present in the SM and PMMM DGMs because of their marginal specification of group means at each visit and unstructured residuals covariance matrix. As a consequence, the random effect for intercept and slope determines the group means (and correlations) [33].

Upon visual inspection we observed the types of trajectories and intercurrent events characteristics as envisaged to be generated, with varying degrees of precision. Acceptability of the variations of precision should depend on the purpose of the simulation.

Any of the four DGMs may be used with different modeling methods, e.g. the PMMM may be used with linear mixed effects models for repeated measures (MMRM) generated patterns instead of marginal model. The probabilities of intercurrent events may also depend on baseline covariates and outcome data. Each of these DGMs may be combined with elements from the other DGMs. For instance, the JM may be combined with a selection rule on the slopes to define trajectories of an intercurrent event more precisely (steep slope leading to AE, positive slope leading to LoE). The α regularization (association) factor had small values in the JM, but the random effects may be given larger weights in the linear predictor of the survival model to create a stronger association.

These DGMs may be used for any other intercurrent event encountered in a trial, e.g. treatment switch, use of rescue medication or liver transplantation. Furthermore, any of these four DGMs can be used to generate fewer or more intercurrent events. Depending on the rules (SM), random effects, fixed effects (intercepts) or survival models used, a patient i could experience two different intercurrent events during the trial at the same or different timing. Different modeling approaches may be used to handle the competing events, if only one intercurrent event is wanted to be simulated per patient. One is to consider only the first of the multiple events, or the most severe or impactful

on the medical condition development if another intercurrent event can exist after the first intercurrent event. For the JM, another approach is to use a Fine-Gray model, or a restriction imposed at patient level (to be not "at risk" for other intercurrent events once an intercurrent event was observed) for the SM or SPM DGMs. For either of the approaches, clinical reasoning and fine-tuning may be applied.

All models need finetuning to establish and preserve the ratios of intercurrent events at trial level. One may need to decide to prioritize AEs over LoE occurring for the same patient (AE earlier). This can be done while obtaining the desired final percentages of LoE, after adjusting for AE first/priority because they are observed earlier at week 2 than LoE are observed at week 3.

Discussion

With this research we propose four DGMs to model the association between outcomes and intercurrent events, illustrated by a target trial in depression. With varying degrees of precision, all four DGMs can successfully simulate the target trials. All deemed acceptable, with the desired longitudinal trajectories and desired types, timings, and percentages of intercurrent events. The SM, SPM and JM are more suited to qualitative replication of the target trial, while PMMM can simulate it with high precision and fidelity.

The Addendum did not explicitly specify or suggest any relationship between intercurrent events and outcomes measured before or after the intercurrent event and the SM, PMMM and SPM/JM implement each different kind of associations.

The SM DGM simulates intercurrent events conditional on outcomes. Rules can be formulated to describe how intercurrent events depend on the longitudinal outcomes. It can be implemented using deterministic rules or stochastic models, both implementations equally successful in simulating intercurrent events. This DGM has an intuitive understanding for lack of efficacy intercurrent events; for instance, if a patient is not recording sufficient efficacy by or at a certain visit, then the patient will experience lack of efficacy. It is possibly the simplest DGM to use from a computational perspective if there are more qualitative and open requirements on the joint distribution of longitudinal data and intercurrent events. In contrast, to reproduce a quite specified target trial may require exploring a range of rules.

The PMMM DGM simulates outcomes conditional on (patterns of) intercurrent events. Namely, the trajectories of patients will depend on the pattern of intercurrent events

they belong to. This DGM also has an intuitive understanding; the patients having a specific type of intercurrent event at a specific timing will likely have some similarity in longitudinal trajectories. It is a difficult DGM to use from a computational perspective. Given data of a source trial, it requires sufficient intercurrent events in each pattern to estimate the longitudinal models of each pattern. This way, it can simulate trials with high accuracy and fidelity to the source trial.

The SPM DGM simulates outcomes and intercurrent events for a patient given her/his associated random effects. Each patient has a propensity to experience an intercurrent event conditional on random effects from the model for longitudinal outcomes (not on the actual outcomes as SM DGM). As the shared random effect models an association, this also holds the other way around: the trajectory before the intercurrent event depends on the intercurrent event via the random effects. The parameters could be estimated from fitting the SPM to the source data if available. It is also complex to use, from a computational perspective, but can simulate intercurrent events in the desired percentages.

The JM DGM also simulates outcomes and intercurrent events that are associated through the shared random effects, but distinctly from the SPM, the generation of intercurrent events is via random effects in the linear predictor of the survival model instead of a logistic model. Also here parameters could be estimated from fitting the model to source data. It can simulate intercurrent events (with timings) distributed throughout the trial and this makes it attractive from a clinical plausibility perspective.

The four proposed DGMs can be used for different objectives. The SM, SPM and JM are less precise DGMs when compared to PMMM, but they are still qualitatively capable of simulating a target trial. If the objective is to have a qualitatively replicated trial, where the aim is to simulate specific patterns of intercurrent events, but without the need to obtain a specific percentage of intercurrent events at trial level, then the SM, SPM and JM are suited. If the objective is to have a precisely replicated trial, where the aim is to simulate specific patterns of intercurrent events, with the need to obtain a specific percentage of intercurrent events at specific timepoints withing each arm, then the PMMM is suited.

Depending on the objective of the simulation, other (trial) characteristics may be needed.

For illustration of the case study, the objective of the simulation was to replicate qualitatively the target trial.

For successful and meaningful use of these proposed four DGMs in simulating trial data with outcomes and intercurrent events, the estimands addendum framework should be used in conjunction. Hence, a multidisciplinary approach is strongly encouraged, and multiple stakeholders should be engaged in the interaction in order to simulate trial data (outcomes and intercurrent events) appropriately. The SM DGM needs and allows the most clinical input for the selection rules to be realistic and plausible, as should be with estimands. The PMMM DGM only needs longitudinal data for each pattern and requires minimal or no finetuning. We found that more work may be needed in the SPM and JM, as refinement may be needed in the logistic or survival model to achieve the targeted frequency (and timing) in the intercurrent events.

The different DGMs differ in their capabilities.

Firstly, we observed in the simulations, that the SPM and JM DGMs encounter more difficulty in capturing the different profiles associated to different intercurrent events than the SM and PMMM DGMs). There is no separation between the longitudinal patterns for the different intercurrent events. They are more mixed together than the ones in SM and PMMM DGMs (See Figures 2-5). Intuitively, this can be understood as follows. Firstly, the shared random effects approach generates longitudinal profiles in the same class (e.g. linear). Secondly, the random effects capture only deviations from the overall mean profile, so the logistic/survival submodels can only select 'bands' of profiles to experience an intercurrent event. Here 'bands' (similar intercept and slope) are random effects' values that in the linear predictor are mapped to the same range. In a real trial, one would expect to observe a kind of separation between patterns, at least to some extent. For instance, one could expect LoE under a "not at random mechanism", but it is simulated using a shared parameter model with a linear-mixed effects model, thus under an "at random" mechanism.

A distinct capability of the PMMM is that the model parameters for treatment effect in each pattern can be refined such that the weighted average of treatment effect at each visit is always as prespecified at (targeted) trial level (See tables 2 and 4).

Furthermore, the SM and PMMM can easily be used to generate other group trajectories than the one used in the case study [33]. If there is a late separation of the treatment effects or if there is an early separation of group trajectories but there is no treatment effect at end of trial, then these two DGMs could be better/easier to use than the SPM or JM.

Another difference between the DGMs is that SPM and JM the intercurrent events can only be associated with observed outcomes, and follow in this sense can only follow an

"at random" mechanism. "Not at random" mechanisms can only be used with the SM and PMMM DGMs and the SM can use actually both (depending on whether the selection rule is based on observed or also not observed data).

When looking at the number of intercurrent evens per patients, the PMMM is the only DGM that can generate exactly one intercurrent event per patient, without competition of multiple intercurrent events on the same patient. This is possible due to the nature of PMMM, namely one pattern is described by one (or more) intercurrent event(s). Unlike the other DGMs, where based on a selection rule or on shared random effects, one patient could turn out to experience intercurrent events coming from different rules or models. Separate handling for these situations would therefore be implemented in close consideration to the rules for intercurrent events (SM) or of the shared random effects (SPM and JM).

The choice between the four proposed DGMs cannot be decided solely on the source data because the DGMs depend on unobserved characteristics of the models and because the source trial data are incomplete. Hence, clinical considerations are important and needed to decide which DGM should be used for simulation studies.

Limitations

We did not evaluate quantitatively how successful the four proposed DGMs were in simulating datasets mirroring the source trial. We relied primarily on visual inspection. We did not define a quantitative criterion.

We did not consider the problem of modeling and simulating intercurrent events and missing data (in the sense of E9(R1), i.e., not following an intercurrent event). We highlight some questions that we did not aim to answer in this research: what if missing data are considered MAR, e.g. depending on outcomes observed up to the missing data, but those outcomes themselves are affected by intercurrent events that are depending on future (unobserved) outcomes (under a "not at random mechanism" such as the SM DGM)? How can this be appropriately modeled?

More research is needed to investigate such questions and the possible implications in other settings or designs beyond the current and specific use of the four proposed DGMs described.

Our proposed DGMs can have various applications.

More than one DGM could be used to generate a target trial. This could be a possible way to verify the robustness of assumptions for that target trial (e.g. at planning and design stages). For instance, for sample size estimation in relation to the constructed estimand for the primary objective. Also, DGMs could be used for evaluation of estimation methods, sample size considerations integrating estimands, alignment of data collection with the targeted estimand, evaluation of multiple estimands of interest that could be estimated from the same trial and other areas such as estimands in adaptive trials, estimands for meta-analysis or estimands in rare disease settings. It could also be the (undiscussed) case that an estimand could be changed based on the results of the simulation, should they indicate so. Although it may discuss whether the targeted estimand should be changed even if the simulations show that the estimation of the estimands may be difficult.

As a final remark, we showed how the four DGMs work based on actual depression trials (source trials). However, these proposed DGMs can be transported to other types of designs, outcomes, intercurrent events, etc. as the idea remains the same.

Conclusion

The four proposed data-generating models can simulate randomised clinical trial data with associated clinical outcomes and intercurrent events. They can be used for simulations to evaluate different estimands and to investigate in-depth their properties. The pattern-mixture mixed model and selection model are more flexible than the shared-parameter models. Understanding and simulating the association between outcomes and intercurrent events may be a solid steppingstone towards successful and proper implementation of the ICH E9(R1) estimands framework for clinical trials planning, design, conduct, analysis, and interpretation.

Acknowledgments

We thank Rutger van den Bor, PhD for double-checking the accuracy and correctness of the R code developed and used for these simulations.

References

1. ICH E9 statistical principles for clinical trials [Internet]. European Medicines Agency. 2018 [cited 2021 May 17]. Available from: https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials
2. ICH Official web site: ICH [Internet]. [cited 2021 May 17]. Available from: https://www.ich.org/page/efficacy-guidelines
3. E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials [Internet]. U.S. Food and Drug Administration. 2021 [cited 2021 May 17]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9r1-statistical-principles-clinical-trials-addendum-estimands-and-sensitivity-analysis-clinical
4. ICH E9(R1) EWG. ICH E9(R1) Estimands and Sensitivity Analysis in Clinical Trials. STEP 4 TECHNICAL DOCUMENT. [Internet]. [cited 2019 Dec 4]. Available from: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf
5. Lasch F, Guizzaro L, Dávila LA, Müller-Vahl K, Koch A. Potential impact of COVID-19 on ongoing clinical trials: a simulation study with the neurological Yale Global Tic Severity Scale based on the CANNA-TICS study. Pharmaceutical Statistics [Internet]. [cited 2021 Mar 10];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.2100
6. Akacha M, Branson J, Bretz F, Dharan B, Gallo P, Gathmann I, et al. Challenges in assessing the impact of the COVID-19 pandemic on the integrity and interpretability of clinical trials. Statistics in Biopharmaceutical Research. 2020 Jul 1;1–18.
7. Degtyarev E, Rufibach K, Shentu Y, Yung G, Casey M, Englert S, et al. Assessing the Impact of COVID-19 on the Clinical Trial Objective and Analysis of Oncology Clinical Trials—Application of the Estimand Framework. Statistics in Biopharmaceutical Research. 2020 Oct 1;12(4):427–37.
8. CHMP. Comirnaty European Public Assessment Report [Internet]. European Medicines Agency. 2020 [cited 2021 Jun 25]. Available from: https://www.ema.europa.eu/en/medicines/human/EPAR/comirnaty
9. Ruberg SJ, Akacha M. Considerations for Evaluating Treatment Effects From Randomized Clinical Trials. Clinical Pharmacology & Therapeutics. 102(6):917–23.
10. Akacha M. Choosing Measures of Treatment Benefit: Estimands and Beyond. CHANCE. 2019 Oct 2;32(4):12–7.
11. Keene ON, Ruberg S, Schacht A, Akacha M, Lawrance R, Berglind A, et al. What matters most? Different stakeholder perspectives on estimands for an invented case study in COPD. Pharm Stat. 2020 Jan 9;
12. Permutt T. Defining treatment effects: A regulatory perspective. Clinical Trials. 2019 Feb 14;174077451983035.
13. Mitroiu M, Oude Rengerink K, Teerenstra S, Pétavy F, Roes KCB. A narrative review of estimands in drug development and regulatory evaluation: old wine in new barrels? Trials. 2020 Dec;21(1):671.
14. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982 Dec;38(4):963–74.
15. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. John Wiley & Sons; 2012. 742 p.
16. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine. 2006;25(24):4279–92.
17. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074–102.
18. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Internet]. [cited 2020 Apr 1]. Available from: URL https://www.R-project.org/.
19. Pedersen TL. patchwork: The Composer of Plots [Internet]. 2020 [cited 2021 Jun 26]. Available from: https://CRAN.R-project.org/package=patchwork
20. Ripley B, Venables B, Bates DM, ca 1998) KH (partial port, ca 1998) AG (partial port, Firth D. MASS: Support Functions and Datasets for Venables and Ripley's MASS [Internet]. 2021 [cited 2021 Jun 26]. Available from: https://CRAN.R-project.org/package=MASS
21. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686.
22. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2019). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-140, <URL: https://CRAN.R-project.org/package=nlme>.

23. Bates D, Maechler M, Bolker [aut B, cre, Walker S, Christensen RHB, et al. lme4: Linear Mixed-Effects Models using "Eigen" and S4 [Internet]. 2021 [cited 2021 Jun 26]. Available from: https://CRAN.R-project.org/package=lme4

24. The janitor package [Internet]. [cited 2021 Jun 26]. Available from: https://garthtarr.github.io/meatR/janitor.html

25. Jr FEH, others with contributions from CD and many. Hmisc: Harrell Miscellaneous [Internet]. 2021 [cited 2021 Jun 26]. Available from: https://CRAN.R-project.org/package=Hmisc

26. Richard Iannone, Joe Cheng and Barret Schloerke (2020). gt: Easily Create Presentation-Ready Display Tables. R package version 0.2.0.5. https://CRAN.R-project.org/package=gt.

27. TheMarianMitroiuTest/Glowing-DGMs-for-outcomes-and-intercurrent events [Internet]. GitHub. [cited 2021 Jun 25]. Available from: https://github.com/TheMarianMitroiuTest/Glowing-DGMs-for-outcomes-and-intercurrent-events

28. García-Hernandez A, Pérez T, Pardo M del C, Rizopoulos D. MMRM vs joint modeling of longitudinal responses and time to study drug discontinuation in clinical trials using a "de jure" estimand. Pharmaceutical Statistics [Internet]. [cited 2020 Aug 2];n/a(n/a). Available from: http://onlinelibrary.wiley.com/doi/abs/10.1002/pst.2045

29. Kenward MG, Rosenkranz GK. Joint Modeling of Outcome, Observation Time, and Missingness. Journal of Biopharmaceutical Statistics. 2011 Feb 28;21(2):252–62.

30. Rizopoulos D, Lesaffre E. Introduction to the special issue on joint modelling techniques. Stat Methods Med Res. 2014 Feb 1;23(1):3–10.

31. Rizopoulos D. JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. ournal of Statistical Software, 35(9), 1-33 [Internet]. 2010 [cited 2021 Jun 26]. Available from: http://www.jstatsoft.org/v35/i09/

32. Drugs@FDA: FDA-Approved Drugs [Internet]. [cited 2020 Apr 1]. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/nda/96/020415Orig1s000rev.pdf

33. Siddiqui O, Hung HMJ, O'Neill R. MMRM vs. LOCF: A Comprehensive Comparison Based on Simulation Study and 25 NDA Datasets. Journal of Biopharmaceutical Statistics. 2009 Feb 25;19(2):227–46.

34. Horrocks J, Heuvel MJ van D. Prediction of pregnancy: a joint model for longitudinal and binary data. Bayesian Analysis. 2009 Sep;4(3):523–38.

Supplementary materials

Code

The R script (verbose) is available for all four DGMs on GitHub at https://github.com/TheMarianMitroiuTest, under a CC-BY-4.0 License.

Appendix 1: Details on the implementation for each DGM

1. Selection model (SM)

a. Stochastic models governing the intercurrent events (IE_i) as a function of outcomes (Y_i)

In the stochastic implementation of the selection model DGM, the intercurrent event IE_{ij} (e.g. treatment discontinuation due to LoE) of patient i at a timepoint j in the trial is Bernoulli distributed with mean p (p = $P\left(\phi\left(Y_{ij}\right)\right)$); p is modeled by a function conditional on the outcomes of patient i at timepoint(s) j in the trial (e.g. J has six timepoints). The used function is a logistic regression model.

$IE_i | Y_i \sim Bernoulli\left(P\left(\phi\left(Y_{ij}\right)\right)\right)$, IE_{ij} being conditional on Y_{ij},

$$logit\left\{P\left(\phi\left(Y_{ij}\right)\right)\right\} = \phi_0 + \phi_1 Y_{i1} + \phi_2 Y_{i2} + \phi_3 Y_{i3} + \phi_4 Y_{i4} + \phi_5 Y_{i5} + \phi_6 Y_{i6}$$

$\phi_j =$ parameters of the linear predictor for Y_{ij}.

For illustration and simplicity, the following model could be used:

$$logit\left\{P\left(\phi(Y_{i3})\right)\right\} = \phi_0 + \phi_1 Y_{i1} + \phi_6 Y_{i6}$$

Or even a simpler model with intuitive interpretability, namely depending on a difference between specific timepoints of the trial, such as from baseline to end of trial.

$$logit\left\{P\left(\phi(Y_{i3})\right)\right\} = \phi_0 + \phi_1\left(Y_{iJ} - Y_{i0}\right)$$

To this logit model, (other) baseline covariates X_{ij} can be added (specified). The logit model can be informed from fitting a model on the available trial data. Alternatively, and depending on the timing and the type of intercurrent event and treatment arm, a logit model can be postulated with ϕ parameters that can be "at random", "completely at random" or "not at random" mechanisms or mixtures of these mechanisms, as aimed for the target trial.

For example, if IE_{i3} (intercurrent event experienced prior to, and recorded for patient *i* at visit *j*=3). Please see the explanation for intercurrent event timing of being experienced and of being recorded.

if $\phi_0, \phi_1, \phi_2, \phi_3 \neq 0, \phi_4, \phi_5, \phi_6 = 0$ then the mechanism corresponds to "at random";

if $\phi_0, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6 \neq 0$, then the mechanism corresponds to "not at random";

if $\phi_0 \neq 0$ and $\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6 = 0$, then the mechanism corresponds to "completely at random" conditional on baseline covariate (could be Y_{i0} for instance);

if $\phi_0, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6 = 0$, then the mechanism corresponds to "completely at random" independent of any outcome or baseline covariate.

For clarity and in all cases presented for the selection model, the outcomes can be observed or not. Observing the outcomes can follow other models formulated for the observation process. This can follow the same or different models and dependencies as for outcomes and intercurrent events.

b. Deterministic rules governing the intercurrent events as a function of outcomes

With a deterministic rule (a function of outcomes), the occurrence of an intercurrent event of a patient can be simulated conditional on the longitudinal outcomes of the patient. IE_{ij} is expressed as a function of a difference in outcome at a timepoint *j* and another outcome for that same patient *i* at another timepoint *j* (if this $j = 0$, then the difference is from baseline); (*j* of the outcome can be different from *j* of the intercurrent event, corresponding to a certain mechanism or dependency). The change (from baseline) is one simple option; in general, it could be any combination of linear (or non-linear) predictors. Thus, it can be of the form:

$$f(\phi_0 + \phi_1 Y_{i1} + \phi_2 Y_{i2} + \phi_3 Y_{i3} + \phi_4 Y_{i4} + \phi_5 Y_{i5} + \phi_6 Y_{i6})$$

In the change from baseline choice, the intercurrent event will occur if the difference in outcomes between timepoint *j* and baseline is bigger or smaller than a certain threshold (δ_{IE}) that fits the clinical assumption or what is observed in the available trial data. For each and any intercurrent event, a different rule with a different threshold can be used to simulate the different intercurrent events.

$IE_{ij} = f(\delta_i),\ Pr(IE) \sim \delta_i$,

$\delta_{ij} = Y_{is} - Y_{ih}$; *s* and *h* belong to *J*

Specification of dependency intercurrent event - outcome

Intercurrent event following an "at random" dependency

$D_{ij} = Y_{is} - Y_{ih}; \; s < h, \; IE_{ij} \sim Y_i^b$

IE_{ij}, where $h \leq j$

Intercurrent event following a "not at random" dependency

$D_{ij} = Y_{is} - Y_{ih}; \; s < h, \; IE_{ij} \sim Y_i^a$

IE_{ij}, where $j < h$

For an intercurrent event (e.g. treatment discontinuation due to lack of efficacy) we can specify a "not at random" dependency ($D_{ij} = Y_{is} - Y_{ih}$; s < h)

IE_{ij}, where $j < h$

$D_{ij} = D_{iJ}$

$D_{iJ} = Y_{ij} - Y_{iJ}$

$D_{IE} \sim q$; q is a pre-specified value (agreed with the clinicians) depending on what is observed a priori in a trial to be used for specific trajectory to reproduce in the simulation.

If $D_{iJ} \leq D_{IE}$, meaning, that the difference in outcomes between timepoints is smaller than the assumed threshold, the patient fits in the prespecified rule, hence the patient i experiences the intercurrent event at $j = J$.

As explained above, the same rule can be defined for a certain intercurrent event to be simulated in both experimental and control arm (e.g. treatment discontinuation due to lack of efficacy), or different rules can be defined per arm (e.g. different rules for treatment discontinuation due to AE per each arm – matching a plausible clinical or pharmacodynamic assumption which would be expected, namely AE in the experimental arm could be due to a possible higher dose (overdose), not possible in placebo, hence plausible under different assumptions). Different rules can also be defined to simulate intercurrent events at different timings, such that the simulated data emulates the target data as envisaged and according to the assumptions made.

In order to control the percentages of intercurrent events, on top of the deterministic rule, a random component (Bernoulli process independent of outcomes and other covariates) can be added, such that there is variability and not all patients that fit in the deterministic rule automatically must experience the intercurrent event. Another possibility to control the percentages of intercurrent events would be to adjust the differences in outcomes such that more or fewer people fit in the rule to simulate an intercurrent event. To this end, the vector of intercurrent events (simulated or not, according to the rule) are then multiplied by a certain probability. The vector of intercurrent events (IE_{ij}) is multiplied by $P \sim Bernoulli(P_{IE})$; P_{IE} is a parameter to be finetuned, for instance depending on what percentages of a certain intercurrent event are wanted to be simulated at trial/arm level.

Supplementary, this P_{IE} can be further sampled from a sample space P containing a sample of $P_{IE}s$ with average P_{IE}. Such that, for instance, in small samples we avoid obtaining the same percentage of that particular intercurrent event at trial/arm level. P_{IE} is sampled for each simulated trial from (range) at each dataset in order to add more variability.

Average P_{IE} is ≈ prespecified $P_{IE'}$ such that the ratios of the intercurrent events within arms/timing are according to parameters envisaged.

This DGM is a useful approach to simulate trials with outcomes and intercurrent events when assumptions can be made to condition the occurrence of intercurrent events on the longitudinal outcomes. Certain rules or models can be used to describe the probability of occurrence of each intercurrent event based on outcomes.

This DGM can implement association between outcomes and intercurrent events under "at random" or "not at random" mechanisms.

"At random" dependency

For an intercurrent event (e.g. treatment discontinuation due to AE) we can specify a MAR-*like* dependency as follows: occurrence at visit J depends on the difference with previous visit $j - 1$.

$$D_{iJ} = Y_{ij} - Y_{iJ}$$

If $D_{iJ} > D_{IE'}$ then the intercurrent event is experienced at $j = J$ for that particular patient i. Here, $D_{IE'}$ is a pre-specified value depending on what is observed a priori in a trial to be used for specific trajectory to reproduce in the simulation.

This DGM approximates a pre-specified percentage of intercurrent events in the generated dataset over all patients according to the rules established on data observed in other trials. To cover a broader range of intercurrent events percentages, a scaling factor can be added to multiply the probabilities of intercurrent events occurrence in the experimental and in the control arms; such that the overall preset percentages are preserved as desired upon multiple iterations of the target trial.

Either of the methods for the models governing the intercurrent events as a function of outcomes can be used, and depending on disease characteristics and patient expected outcomes, incorporating available knowledge (e.g. available phase II or III trial data).

This DGM can simulate linear trajectories, or more complex and different scenarios, such as no effect during trial with (very) late separation of trajectories at the end of trial, early separation and no effect at the end of trial, or other different trajectories [33]. This DGM can simulate intercurrent events at only one timepoint/visit with one model. If intercurrent events are aimed to be experienced at multiple timepoints, then multiple models are needed, hence more assumptions (identified) or knowledge needed (from other trials).

2. Pattern-mixture mixed model (PMMM)

Pattern models can be derived directly by postulating the longitudinal outcome trajectories and assigning at which timepoint they will experience the specific intercurrent events (corresponding to the pattern they belong to) according to clinical assumptions based on observed trial data. Another implementation, more suited if trial data are available, could be to fit pattern-mixture models on the trial data for the intercurrent events of interest (type, timing), and then directly use these models to simulate trial data.

Depending on the desired implementation, we can add a sampling function (e.g. a range of values, via a uniform distribution centered at the value of proportion for each or some pattern $U\left[\text{proportion}_{\text{k}} - \text{margin}_{\text{k}} \ : \text{proportion}_{\text{k}} + \text{margin}_{\text{k}}\right]$, with mean equal to the proportion for pattern k) to vary the number of patients from each pattern k for each iteration of a trial, while maintaining the desired percentages of each pattern (or each intercurrent event) at specific values over all iterations. This approach can be omitted, if the aim is to generate exactly the same constant number of patients in the corresponding patterns in simulated trials (e.g. q patients in pattern one and v patients in pattern two kept constant in all trials). If this approach is used, we can add ranges of probabilities for each pattern when generating each trial data, and hence we will have a variable number of patients each time in each pattern in each trial (centered around

the mean = proportion_k), keeping the distribution of patients over the patterns within a desired range. This must be reflected in the model parameters for each pattern, such that at trial level (upon stacking all patients from all patterns) the parameters are in concordance with the target trial parameters (e.g. treatment effect at end of trial). After generating these patterns of patients experiencing intercurrent events, we can generate the (remaining Sample size – patients in all other patterns) patients not experiencing any intercurrent event. This can be done by randomly allocating these patients to the trial arms or by specifying the exact number of patients in each arm.

Depending on the percentages of intercurrent events wanted to be generated, we use can use $P(IE|\ k_{pattern})$ to generate n of these patients for each pattern k. We use specific proportions and ratios (constant across simulated trials) in our case study for simplicity and to have more control on what percentages exactly we want to generate, as well as to ensure comparability of simulated targeted trials with the other three proposed DGMs.

Hence, we have a specific proportion of each pattern, each pattern with its own model:

Table 9. Pattern-mixture mixed model specification and notation

Pattern	Model specification
No intercurrent events	$Y_{ij}^{NoIE} = \left(\beta_{1j}^{NoIE} + \varepsilon_{ij}^{NoIE}\right) + \beta_{2jTreat}^{NoIE} x Treatment_i$
Treatment discontinuation due to Intercurrent event (IE) IE could be lack of efficacy (LoE), adverse event (AE), death, etc.	$Y_{ij}^{IE} = \beta_{1j}^{IE} + \varepsilon_{ij}^{IE} + \beta_{2jTreat}^{IE} x\ Treatment_i$ If by arm: $IE_{Treat}: Y_{ij}^{IE_{Treat}} = \beta_{1j}^{IE_{Treat}} + \varepsilon_{ij}^{IE_{Treat}} + \beta_{2jTreat}^{IE_{Treat}} x\ Treatment_i$ $IE_{ctrl}: Y_{ij}^{IE_{ctrl}} = \beta_{1j}^{IE_{ctrl}} + \varepsilon_{ij}^{IE_{ctrl}}$
$\varepsilon_{ij}^{x} \sim N\left(0, \Sigma_{\varepsilon_{ij}^{x}}^2\right)$, ε_{ij}^{AE} could have a different or the same covariance matrix as ε_{ij}^{LoE}	

Parameter interpretation is the same as described in Table 1, specifically for each pattern.

Subscript i and j indicate the patient and visit, superscript indicates the type of intercurrent event and the arm for which to be simulated. $Y_{ij}^{LoE_{Treat}}$ is the outcome for patient i at timepoint j, for which lack of efficacy is simulated in the treatment arm. Hence, we use a mixture of patterns that will comprise the full data Z (Y_{ij} and IE_{ij}) generated concomitantly (outcomes and intercurrent events with one model for each pattern). This could also be implemented via a linear mixed model, instead of the marginal model. Each pattern is simulated with marginal models with unstructured covariances different for each pattern.

This approach can be further extended if needed. For instance, one can define patterns with some or all measurements available after the intercurrent event timing at visit *j*, or simply no measurements at all after the intercurrent event is experienced and recorded, should a terminal event (e.g. death) be the simulated intercurrent event. In our case study. we consider all outcomes and intercurrent events observable (e.g. no deaths).

The PMMM DGM requires trial level parameters, arm level parameters, pattern level parameters according to postulated models for each pattern, proportions of each pattern and timings for each intercurrent event. In other words, the patterns can depend on arm, timing or other characteristics of the intercurrent events. It requires in-depth prior knowledge and accurate assumptions for each pattern in order to be implemented successfully. It is very flexible due to the saturated marginal model (MMRM) which is able to model timepoint means at group level, and can allow for any trajectory of means to be simulated [33]. Also, because in theory, each patient can be a specific pattern corresponding to an intercurrent event, hence it can generate as many patterns as are needed, useful for instance in rare and ultra-rare disease settings. It can also simulate (perfectly) linear trajectories or more complex and/or different scenarios (similarly to DGM **Selection Model**), such as no effect during the trial with late separation of trajectories at the end of the trial, early separation and no effect at the end of trial [33] or other possible trajectories. It can simulate with one model intercurrent events at one or more timepoints. If intercurrent events are aimed to be experienced at multiple timepoints, the same models but also multiple or differentiated models could be used, hence more assumptions or knowledge would be needed. Different models can be used for each intercurrent event by arm, but it is operationally simpler to use fewer models, hence, if possible and if the model can capture both patterns, a parsimonious DGM is preferred and recommended (e.g. use one model to describe AEs in both arms). This is strictly dependent on the assumptions made at the level of each pattern and the source trial data. The decision to apply either of the implementations is dictated by how well one or more models can capture the different patterns. The PMMM can be simple, but it can quickly become complex with non-parsimonious models.

This DGM is a useful approach to simulate trials with outcomes and intercurrent events when assumptions can be made to condition the longitudinal outcomes on the occurrence of intercurrent events. Different patterns can be formulated and for each pattern a longitudinal trajectory is assumed. For instance, in the LoE pattern the trajectory will be upwards to denote lack of improvement.

This DGM can implement association between outcomes and intercurrent events under "not at random" mechanisms.

3. Shared-parameter model (SPM)

To generate certain percentages of intercurrent events one can steer them directly from the logit models. One way would be to interpolate the intercept from a certain domain of values, that will ensure a certain range (approximate or precise) of intercurrent events percentages of, while possible keeping other parameters in the model specification constant (e.g. treatment parameter). However, this step may require finetuning that may not be obvious immediately. A simple solution is presented in the available R code [27] with a function that derives and interpolates values of the intercept upon multiple iterations and identifies the values that result in desired percentages of intercurrent events. This corresponds to algebraically equating the intercept from the linear predictor of the logit function of the desired probability. For simplicity and illustration of how this DGM works, we used a fixed effects logistic regression model fitted on a large, simulated trial (using SM DGM), with treatment as covariate for LoE and just with an intercept for AE. To these estimated parameters, we added the (subject-specific) random slopes from the longitudinal model in the logit model (used to generate the intercurrent events) that gives the now subject-specific probabilities to experience the intercurrent event. Other approaches could be employed to use the shared random effects or to extract them from other trials, with specific limitations [34]. A generalized linear mixed effects model could be fitted on the source trial to extract the random effects characteristics (Variance), understand their weight and finetune the α regularization factor in the logit models specification. Depending on the concordance wanted between the source trial and target trial, finetuning of parameters will be necessary. If different percentages of intercurrent events are targeted, then the intercept has to be determined to match the percentage targeted. This may be very different from the intercept from the model fitted on the source trial, simply because the percentages (in the large) are different even if the longitudinal outcomes are then unchanged.

Another implementation to generate certain percentages of intercurrent events, would be with a Bernoulli process that can be employed to multiply the indicator variable of the intercurrent events (resulting from the logit models with random effects). This approach could be used when less information is available and certain model parameters and trial level parameters could be postulated. For instance, with available data from a non-randomised trial. Furthermore, the decision to include or not a random effect or the weight it is given, must be anchored in a clinical assumption. An AE could depend more on the steep improvement of outcome (slope) after treatment initiation, than on the outcome at baseline (intercept). Hence random slope could be specified in the model, while the random intercept could be omitted or given a lower weight. Similarly, LoE could depend on improvement in time of the outcome (or lack thereof) and on the outcome value at baseline. Hence, both random intercept and random slope could be in the model. With

different weights, then it may depend more on how the patient starts feeling in the trial, and less on the starting point (0.1 x random intercept + 1 x random slope, for instance). Other clinical assumptions, intercurrent events and association relation with the random effects can be envisaged.

It can simulate with one model intercurrent events at one timepoints. If intercurrent events are aimed to be experienced at multiple timepoints, different models but also multiple or differentiated models could be used, hence more assumptions or knowledge would be needed.

The sharing of random effects implies limitations on the combinations of trajectories and intercurrent events that occur. The random effects capture a latent propensity for intercurrent events, and all patients with similar random effects or similar combination of random effects (i.e., longitudinal profile) will have similar probability for that intercurrent event. Vice versa, the patients experiencing the same intercurrent events will be homogeneous in their time profiles. This contrasts with what DGMs **Selection Model** and **Pattern-mixture mixed model** can simulate. This (simple) SPM DGM can simulate intercurrent events at only one timepoint with one model. If intercurrent events are aimed to be experienced at multiple timepoints, the same models but also multiple or differentiated models could be used, hence more assumptions or knowledge would be needed.

This DGM is a useful approach to simulate trials with outcomes and intercurrent events when assumptions can be made to condition the longitudinal outcomes and the occurrence of intercurrent events on latent (unobserved) factors (random effects). Different strengths of association can be formulated, and different clinical assumptions can be reflected in the model specification of random effects. For instance, the patients with positive random slopes are more prone to experience LoE.

This DGM can implement association between outcomes and intercurrent events under "at random" mechanisms.

4. Joint modeling of repeated measurements and intercurrent events via JM

To reach the Weibull distribution that best describes a certain intercurrent event in a trial, a refinement of the parameters is needed, process that can be lengthy. However, it is a decisive step in order to simulate trials as targeted.

A possible implementation to obtain targeted percentages of intercurrent events, is to standardize the resulting times to intercurrent events to fit beyond the trial duration. Hence some patients will experience the intercurrent events during the trial, while others will experience it beyond end of trial, therefore, not experiencing during the trial. This implementation can be used to generate different percentages of intercurrent events. The standardization can take into account the quantile values corresponding to certain percentages of intercurrent events. Similarly with SPM DGM, another implementation to generate certain percentages of intercurrent events, would be with a Bernoulli process that can be employed to multiply the vector of survival times of the intercurrent events (resulting from the survival models with random effects). This approach could be used, for instance, when less information is available and certain model parameters and trial level parameters could be postulated.

For both implementations, the difficulty lies in finetuning the parameters of the Weibull distribution, such that intercurrent events are distributed as desired during the trial duration (e.g. 50% of the intercurrent event at visit 2, 30% at visit 3, 15% at visit 4 and 5% at visit 5 and 6). This particular step can be informed by fitting a parameterized Weibull model on available (real) trial data. If less knowledge is available, a simpler (with fewer parameters) exponential distribution can be used

This DGM can be used, for instance, when follow-up is for a longer period of time and using different models for intercurrent events at specific fixed time points is not practical or not desired, or accurate information not available or practical. The shared parameter influences the intercurrent events timing through their specification in the linear predictor of the survival model. This DGM can generate intercurrent events for a range of visits (survival model), not only at fixed specific timepoints as the other three DGMs proposed.

Depending on the prior knowledge of the trial to be simulated, this DGM can be also implemented using a selection rule based, for instance, on the slopes. E.g. the patients with a positive slope (no efficacy or disease worsening) would belong in the LoE group. This approach can be used to steer the trajectories specifically for each pattern. In this case also, a Bernoulli process can be employed to vary the percentages of intercurrent events. For the JM DGM application, the Weibull distribution allows for some (possibly limited) flexibility.

With the use of the linear predictor, there is control on the timing of intercurrent events, but there is no control on the (specific) trajectories the patients (experiencing the intercurrent events) have. Hence, there may be some patients that are doing well clinically (based on the longitudinal outcomes), but they will experience the intercurrent events early/late because the survival function imposes so based on their combination of linear predictors (including random effects). With this DGM, intercurrent events can be generated, but this choice may not be too flexible, compared for instance with the selection model DGM or pattern-mixture mixed model DGMs. This DGM can simulate with one model intercurrent events at different timepoints distributed throughout the trial.

This DGM (JM) could be used when less/limited prior knowledge is available or when the trialists do not want to impose too many restrictions on the simulation, e.g. by postulating precise patterns for a specific intercurrent event.

As this DGM is also a type of shared parameter model, it shares similar properties as Shared-parameter model DGM (Please see SPM DGM details about random effects inclusion in the model). Unlike Selection Model, Pattern-Mixture Mixed Model and Shared-Parameter model DGMs, this DGM can easily simulate with only one model intercurrent events at a range of visits due to the survival model used in the intercurrent events sub-model. If more specific trajectories are wanted to be simulated, a selection on the slopes could be implemented, for instance, to define the patients' trajectories in relation to an intercurrent event (e.g. patients with a negative slope to experience treatment discontinuation due to adverse events, patients with a positive slope to experience treatment discontinuation due to lack of efficacy).

When simulating longitudinal outcome data, there could be values that are generated and are above or below the natural scale of an outcome observed in a clinical setting. We did not set them to certain values in our simulations. If unusual trajectories are observed, then a check of the simulated data needs to be performed to make sure that simulated data generates the target trial data accordingly. However, if outcomes above or below natural limits occur, these can be truncated to min/max of the values depending on the extent of occurrence of such values. A check of the extent of potential bias introduced systematically needs to be conducted (e.g. comparison of performance measures based on datasets with rounding vs without rounding). Otherwise, another distribution than multivariate normal could be assumed.

We checked to see how often such values were generated using the postulated models and occurrences were very few with minimal impact on parameters.

This DGM is a useful approach to simulate trials with outcomes and intercurrent events when assumptions can be made to condition the longitudinal outcomes and the occurrence of intercurrent events on latent (unobserved) factors (random effects) to simulate intercurrent events throughout the trial with different distributions. Different strengths of association can be formulated, and different clinical assumptions can be reflected in the model specification of random effects in the survival model.

This DGM can implement association between outcomes and intercurrent events under "at random" mechanisms.

Appendix 2: Verification of simulations

We performed the verification of the simulated data (outcomes and intercurrent events) to ensure that what was intended to be simulated (according to the DGM set of parameters) is mirrored by the actual simulations in practice (parameters of simulated datasets).

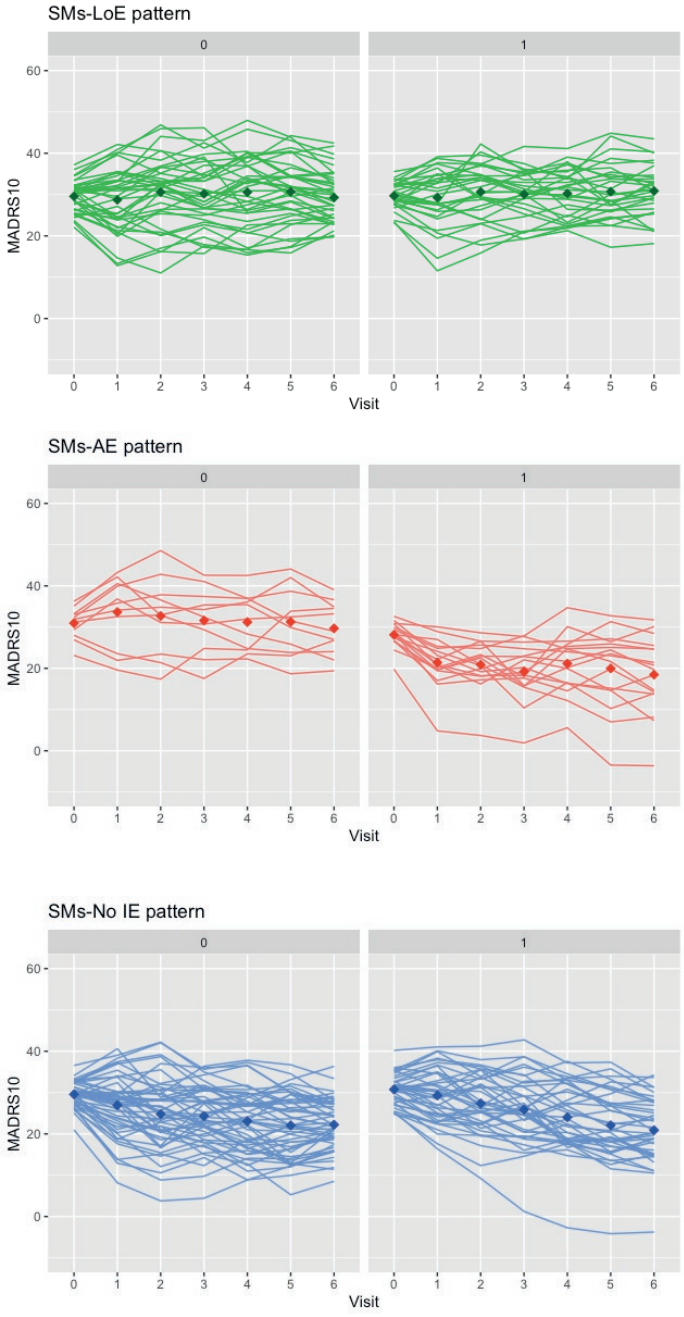
The longitudinal outcomes were verified by retrieving the DGM parameters. The means of estimated treatment effects from the trials generated, were within the tolerance margin, for each of the four proposed DGMs. The intercurrent events were verified by comparing the tabulated aggregate descriptive statistics of the intercurrent events type and percentages, with the DGM parameters, namely the parameters from the models governing the type, timing and percentages of intercurrent events. The summary statistics for the percentages are according to the desired percentages, for each intercurrent event with corresponding timing, at trial and arm level (See tables 8a-d).

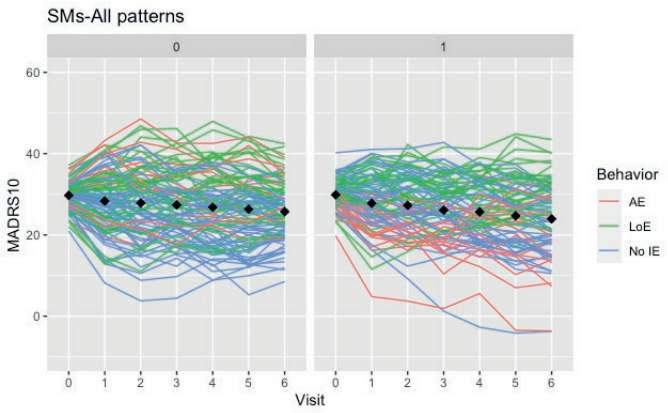
Appendix 3: Verification of longitudinal outcomes

For the verification of longitudinal outcomes, we used a tolerance margin for precision of 0.1 from the true value of the treatment effect at end of trial at six weeks, and derived the number of trials needed for a 95% CI [16]. $z_{(1-\alpha/2)} \times \frac{sd}{\sqrt{n_{trials}}} < 0.1$, plugging in the parameters (z-score and SD), following rounding to an integer it results that we need to simulate 382 for SM, 416 for SPM/JM and 482 trials for PMMM (based on model estimated error of the treatment effect estimate). We chose to simulate 500 trials. The mean of treatment effects estimated from the n_{trials} must be within 0.1 difference of the true treatment effect in order to conclude that the verification of the longitudinal outcomes was successful.

Appendix 4: Graphs for SM DGM stochastic (SMs) implementation

SM stochastic approach





Percentages of intercurrent events for SMs DGM

Table 8e. Descriptive statistics intercurrent events

Selection model DGM - stochastic

	Control		Treatment		Total	
Intercurrent event	N	%	N	%	N	%
AE	10.7	5.6	18.9	9.9	29.6	15.6
LoE	38.8	20.4	26.8	14.1	65.6	34.5
Total	49.5	26.0	45.7	24.0	95.2	50.1

Averaged over 500 simulated trials. Trial sample size = 190

CHAPTER 6

General Discussion

Randomised clinical trials conducted to generate confirmatory evidence of the efficacy of experimental treatments are expected to be used as the basis for scientific evaluation and regulatory approval [1]. Their purpose is to provide robust information regarding the treatment effects of medicinal products in populations of patients diagnosed to suffer from a certain disease. However, estimation and interpretation of treatment effects can be complicated by the occurrence of withdrawals from the study, administration of rescue or prohibited medication, treatment switches, treatment discontinuations due to adverse events or lack of efficacy, or other related events [2–4]. Estimating treatment effects is even more challenging when such treatment effects are not well defined at the trial design stage, and data collection and estimation are not aligned to estimate the targeted treatment effects [5,6]. Due to imprecision in defining which question the trial should address, healthcare stakeholders may lack understanding of the estimated treatment effects. The Intention-To-Treat and Per Protocol analyses were the established analyses for a long time in drug development for treatment effect estimation [4,7–12]. Other analyses have become increasingly popular [3,5,9,13–24]. But why have these analyses become popular, and was there a need to change the norm? The Intention-To-Treat analysis does not fit all scientific questions of interest and may be challenging to apply. Selection bias and confounding may be introduced in a Per Protocol analysis, and the target population cannot be defined outside the trial [13,15].

To address these issues, the ICH E9(R1) addendum on estimands provides a framework to address clinical questions targeted in clinical trials by defining (more) precise treatment effects, aligning the design and analysis with them, and ensuring harmonised communication between involved parties [9,14]. The estimand is a precise description of the treatment effect reflecting the clinical question posed by the trial objective [25]. An essential attribute of the estimand is represented by the strategies to handle intercurrent events. Intercurrent events occur after treatment initiation and affect either the interpretation or the existence of the measurements associated with the clinical question of interest. The events mentioned above, such as study withdrawal, treatment discontinuation and administration of rescue medication are examples of intercurrent events.

The estimands framework is increasingly being used by drug developers and regulators when designing trials, discussing protocols, writing statistical analysis plans, and following the conduct of the trial, also in writing clinical study reports [16–18,21,26–28]. However, it is still to be understood what the added value of this framework is for drug development. Does it indeed lead to improvement of clinical trials and more precise formulation of clinical questions? And does it facilitate understanding of the targeted treatment effects and what is needed to use it to its full potential to benefit patients?

This thesis: Summary

In this thesis, I investigated the (implicit) use of estimands in the pre-estimand era in clinical trials for regulatory evaluation and approval to understand what estimands may correspond to common efficacy analyses. Additionally, to understand the characteristics of estimands more fundamentally, I developed data-generating models (DGMs) that take into account the association between clinical outcomes and intercurrent events.

In **Chapter 1**, I refer to the first randomised controlled trial reported in the scientific literature and explain the randomisation principle. I summarise the development of the ICH E9(R1) estimands framework and the estimand concept, and I present the objectives and the outline of this thesis.

In **Chapter 2**, I promote broadly the need to more clearly define treatment effects in clinical research. I constructively critique that Intention-To-Treat and Per Protocol analyses do not answer all relevant scientific questions (unbiasedly) and bring to the attention of the healthcare stakeholders that the estimands framework encourages targeting and defining treatment effects beyond Intention-To-Treat and Per Protocol analyses. I advocate a continued discussion in the scientific community to continue learning and ensure that the estimands framework implementation improves clinical trial practices.

In **Chapter 3,** I evaluated past estimation practices in drug development and evaluation, viewed from the estimand perspective. The five attributes (treatment, population, variable, population-level summary, and strategies for other intercurrent events not captured yet at the level of mentioned attributes) were present in clinical trials before the introduction of the estimands framework but not in an explicit fashion. I reverse-engineered the estimands from EMA clinical efficacy guidelines, clinical study reports, and statistical analysis plans submitted by the sponsor to support the marketing authorization application and from the list of questions of the regulators. The novelty the estimands framework introduces is the precise description of treatment effects and the explicit use of strategies for intercurrent events integrated into the primary outcome or estimation.

Extracting and deriving the intercurrent events was difficult because relevant information pertaining to estimand attributes was scattered throughout the clinical study reports and statistical analysis plans, heterogeneously between clinical study reports and between medicinal product developments. Strategies for intercurrent events could not be derived from all regulatory questions. The treatment policy strategy was most

often advised in EMA clinical efficacy guidelines, whereas the hypothetical strategy was the leading strategy applied in sponsor documentation, despite not being advised in any disease guideline. Results indicate that the regulatory target of estimation and what is actually estimated are not fully concordant, primarily due to limitations in the collection of outcomes data to enable a treatment policy strategy for all intercurrent events.

The research for this chapter showed the need to start with a precise description of targeted treatment effects and align trial design and analysis. To this end, further dissemination, training, and uptake of estimands in clinical efficacy guidelines based on multidisciplinary discussions are needed to more precisely define a meaningful target of estimation for different stakeholders. It also showed that when outcomes are missing, although treatment policy may be the target, the use of treatment policy is not possible in the strict sense, and this needs to be explicitly addressed.

In Chapter 3**,** I looked from an estimand perspective at estimation practices in drug development and evaluation in written form (EMA clinical efficacy guidelines, Sponsor's protocols, statistical analysis plans and clinical study reports, and regulatory questions during scientific evaluation). In **Chapter 4**, I looked from an estimand perspective at common analysis methods using concrete patient trial data. I re-analysed six randomised controlled trials evaluating a new anti-depression treatment in adults to understand what estimands correspond to common efficacy analyses and to assess the difference in the size of the estimate of the different analysis methods used. I selected the following analysis methods: Analysis of covariance (ANCOVA) on complete cases, ANCOVA following last observation carried forward (LOCF) imputation or multiple imputation, mixed-model for repeated measures (MMRM), MMRM following LOCF imputation or MMRM following jump-to-reference imputation, and pattern-mixture mixed models. I also included a principal stratum analysis on a stratum (of interest) of the study population that would not discontinue due to adverse events or lack of efficacy. I translated each analysis into the targeted estimands and the corresponding clinical questions.

I found that the same analysis method could be matched to more than one estimand, and the same estimand could be matched to more than one analysis. The major differences between estimands concern the choice of strategy for each intercurrent event.

Not all six estimands had a clinically meaningful interpretation. This suggests the need to formulate estimands in clinical trials in a multidisciplinary team, especially involving clinicians and statisticians. Imputations are often used in conjunction with analysis methods without consideration of the alignment of the values imputed with the trial objective. Another key finding is that no distinction is made between the reasons that

led to missing data, such as lack of efficacy or an adverse event. Both are usually imputed in the same way, while in terms of clinical interpretation, the two deserve different considerations in determining any imputation strategy.

Only a few analyses would target the same estimand; hence, by definition, few could be used as sensitivity analyses. Some analyses used and communicated as sensitivity analyses were not truly sensitivity analyses according to ICH E9(R1), as they target a different estimand rather than testing the sensitivity of assumptions for the estimation of the same estimand. Some analyses targeted the same estimand, but they were not sensitivity analyses of each other as they had the same underlying assumptions. For example, ANCOVA following MAR multiple imputation and MMRM rely on the same assumptions.

Our findings emphasize that estimands should be prespecified because when retro-fitting estimands, there is no 1-to-1 mapping between common analysis methods and estimands.

In **Chapter 5,** I aimed to develop and evaluate data-generating models to jointly simulate outcomes and intercurrent events for randomised clinical trials. I proposed four DGMs for the joint distribution of longitudinal continuous clinical outcomes and intercurrent events under the scenario where they are observable (meaning no terminal events occurred during the observation period): a selection model, a pattern-mixture mixed model, a shared-parameter model, and a joint model of longitudinally observed clinical outcomes and a survival model for intercurrent events. I used a case study in a short-term depression trial with repeated measurements of continuous outcomes and two types of intercurrent events (lack of efficacy and adverse event) and compared the four proposed DGMs. I proposed possible ways to implement these DGMs. I found that all four DGMs can mimic a target trial in terms of envisaged patterns of intercurrent events and outcomes trajectories with varying degrees of precision. Prior available clinical trial data on which models and parameters can be based can be used to reach realistic target trials.

The four proposed DGMs can simulate a broad range of scenarios that can enable the evaluation of different estimands and allow in-depth investigation of their properties. Consequently, simulation studies using these DGMs can increase fundamental understanding of estimands and can inform planning, design, conduct, and analysis of randomised clinical trials, as well as interpretation of the results trials generate.

Implication for practice

Patients and other healthcare professionals

I bring to the attention of the healthcare community that there is a new scientific guideline to better describe and estimate treatment effects in clinical trials. I also advocate that the currently often used and reported treatment effects are not the only treatment effects of interest.

The estimands framework opens new opportunities in drug development and can be used as an opportunity for patients to shape their own critical role in clinical trials, to be more involved in decision-making at all stages of new treatment development. Patients' opinions can be better integrated into the trials conducted for the diseases they are affected by or in trials in which they participate. There are more estimands, and some may be more interesting for patients while not being of primary interest to other stakeholders, such as regulators or medicines reimbursement agencies. It is unclear what is the value of Intention-To-Treat or Per Protocol analyses for patients; possibly, there are other treatment effects of more interest for patients. This would not necessarily translate to mutual exclusivity of the different and possibly many estimands of interest, for the different stakeholders, in clinical trials.

Patients are in a position to drive trials on paths of estimating treatment effects that are more meaningful since they would know best what works and what matters for them. For instance, via principal stratum strategy, patients could drive the clinical questions and corresponding estimands formulation much more dominantly because there are more options created and made available, and because they could be involved in this debate more prominently and from the beginning, from trial objective formulation and planning.

Other adjacent opportunities are also opened, e.g. for Patient-Reported Outcomes (PRO) [29] which is an area that could benefit from integration with the E9(R1) estimands framework [30].

Biostatisticians and Clinicians

The estimands are a great responsibility to be shared between biostatisticians and clinicians involved from the initial stages of trials, from conduct and analysis to interpretation of results. The DGMs proposed in this thesis are a novel way to investigate estimands in-depth with simulation studies. They provide control on the association between outcomes and intercurrent events, and can be used to evaluate the properties of various possible estimands of interest for different stakeholders. They can be informed by other

trials and by clinical subject-matter expert knowledge. Multidisciplinary collaboration and genuine interactions need to be ensured to align trial objectives, data collection, analyses, interpretation and communication of trial results, employing the estimands framework at every stage.

Medicine regulators and drug reimbursement agencies, as organisations

I advocate the estimands framework, and I constructively criticize Intention-To-Treat and Per Protocol analyses. While they were useful before the estimands framework, it became obvious they are either not fit or not the only treatment effects of interest in clinical trials and for all stakeholders.

Although, some implicit estimands thinking is already present in regulatory guidance, I envisage a general and complete revamp of clinical efficacy guidelines [31,32] and an improved methodology, with explicit use of estimands addendum, for clinical trials supporting marketing authorization applications. However, the application and integration of the estimands framework in drug development are still in the beginning phase. It may not be easy, nor is it likely to happen fast. Still, certainly, it is needed to ensure alignment with the ICH E9(R1) addendum to facilitate estimands-driven drug development and regulatory evaluation.

It is not uncommon that different regulatory agencies can provide scientific advice or protocol assistance with different recommendations to sponsors who intend to conduct a certain trial. Perhaps by improving the precision of targeted treatment effects, the estimands framework may not necessarily lead to more or less harmonised advice. Still, in any case, it should lead to adding more clarity for differing or converging recommendations and justification for the different choices made and recommendations should be provided to reach a final advice.

I envisage the provision of scientific advice and protocol assistance to continue to become much improved and clearer, provided the requests for regulatory advice contain detailed, estimands-supported, and DGMs-informed questions. Possibly, this can ensure consistency between the estimands advised by the regulatory agencies and estimands used by sponsors. The proposed DGMs can also be used to evaluate multiple estimands of interest for the same trial(s). For instance, some estimands to be used for efficacy evaluation and regulatory approval, other estimands to be used for reimbursement and pricing, and one or more estimands to be used for the therapeutic decisions of prescribing physicians. There is considerable research published for estimands to be used in clinical trials, but the focus is on efficacy; there is less or limited attention yet

on this dimension of estimands possible use and added value. Furthermore, there is no formal stance taken by reimbursement agencies or HTA organisations regarding the application and possible added value of the estimands framework [33]. This is uncharted territory that could be mapped with the use of the proposed DGMs and case studies.

Furthermore, I found that information pertaining to estimand attributes is scattered throughout the protocols, statistical analysis plans, and clinical study reports. Our research can contribute to the current pool of knowledge as further basis to drive the current ongoing initiatives to integrate the estimands framework in relevant key templates (statistical analysis plans, protocols, synopses, clinical study reports [34,35]). This wave of infrastructural changes to integrate estimands framework in all dimensions of drug development is already happening. However, possibly due to the setting imperative (scientific evaluation and regulatory approval), it is likely reaching regulatory agencies and industry quickly, but it may not reach academia at the same time. There are some challenges, but they could be overcome by a multidisciplinary and multistakeholder approach and collaboration. This is happening, but not nearly to the full possible extent.

There is still a gap between pre-approval and post-approval, each stage with different challenges and opportunities, for which the scientific community needs to engage in collaborative learning.

Avenues for future estimands research and implementation

The implementation of the estimands framework and the understanding of the properties and selection of estimands is still in its infancy. Ample opportunities exist and are needed for the estimands framework to mature and become a standard element of all aspects in clinical drug development. What avenues are there, and which one should be taken first?

The estimands framework deployment involves change management (training, mindset in clinical trial practice) constituting a human process and an infrastructural change. Furthermore, many other essential topics in drug development were not explicitly discussed by the ICH E9(R1) estimands framework but are impacted by it. For instance, sample size estimation and handling multiplicity, including multiple estimands for multiple stakeholders, and meta-analysis (especially when combining different estimands or trials designed using the estimands framework with older trials). Also, it is not clear yet which estimands should be formulated to evaluate non-inferiority and

(bio)equivalence/similarity to replace the current norm evaluating both the Intention-to-Treat and Per Protocol results. This could be addressed and started to be investigated with case studies and continued with simulation studies.

The estimands framework requires time and attention to become standard practice in drug development. Estimands training must reach all trialists in all environments, academia, pharmaceutical industry, and regulatory bodies to facilitate this process [22]. To do so, estimands need to be made tangible. Therefore, published case studies of development programs with estimands used in the development and regulatory evaluation are required. They also should find a place in reporting guidelines, such as the CONSORT statement [36,37] or in risk-of-bias assessment tools [38]. I discuss below possible suggestions for change management by major stakeholders.

Academic environment

Researchers from academia could identify setting-specific problems, e.g. certain intercurrent events that may occur or may be of interest only for a particular setting. This could highlight yet unknown challenges, for example, handling two different intercurrent events being experienced simultaneously. Or need of further refinement of the estimand, accounting for a series of intercurrent events with a specific sequence that may have different clinical meaning than when singularly experienced. This process of "retro-fitting" estimands is needed because before diving into formulating new estimands, possibly the currently used estimands may be good enough, but they need to be disentangled and made more explicit.

Regulatory network

A critical infrastructural change that could help the adoption and implementation of estimands in trials would be to (continue to) update the disease and methodological guidelines with recommendations of the estimands framework.

Regulatory Agencies (e.g. EMA, FDA, etc.) could set up and co-lead taskforces together with researchers to update the EMA clinical efficacy guidelines [31] according to a 3-5-year workplan. To achieve this objective, they could ensure the environment for a consortium or multidisciplinary groups (statisticians, prescribers, clinicians, patients, and principal investigators). Estimands should be mapped for diseases and corresponding research published (possibly accompanied by reflection papers) to drive and facilitate updates of clinical efficacy guidelines. Academia can play an important role in defining and developing meaningful estimands, and further advance the application of the estimands framework. In oncology, OS, PFS, EFS, and DFS use and assessment will very likely remain unchanged, but the estimand attributes could be made more obvious, for

instance, the treatment attribute or the strategies for addressing intercurrent events [39]. In HIV trials, the established snapshot algorithm might still be the main thinking for treatment effects for obvious reasons, such as ensuring consistency between trials.

Also, they could deliver training to academia, trial sponsors, and clinicians. Furthermore, for the methodological guidelines, such as "Guideline on the choice of the non-inferiority margin", task forces co-led by CHMP working parties, such as the Methodology Working Party, should also be set up to update them.

The assessment reports templates (Day 80 AR, Day 120 List of Questions [40]) should also be updated to accommodate the estimands framework. This is of crucial importance and may not be facile to implement immediately as the actual questions to the sponsors should follow the estimands framework beyond specific sections added in the document templates. The European public assessment reports (e.g. EPAR [41]) and summary of product characteristics (SmPC [42]) templates should also be updated to accommodate for estimands in communicating trial and evaluation results to patients, carers, and prescribers.

Scientific advice and protocol assistance procedures should make estimands questions mandatory and addressed with top priority.

Pharmaceutical industry

They are one of the leading clinical research engines and main initiators of clinical trials for clinical development plans. Pharmaceutical companies could be in a great position, together with academia (approximately 33% of all phase I-III studies are funded by academia [43]), early in clinical development to ensure trials address meaningful clinical questions, for patients, for regulators, and reimbursement agencies. When defining objectives for proposed clinical trials and discussing the designs to address the formulated objectives, pharmaceutical companies should involve patients, regulators and reimbursement agencies early. Sponsors in the pharmaceutical industry lead several initiatives for estimands for specific interest groups, such as oncology or neurology [44,45], which are much needed. More such multidisciplinary cross-company initiatives should be set up and conducted in other disease areas to spearhead the estimands research. This could possibly contribute to the reflection papers needed to update clinical efficacy guidelines (mentioned above in "*Regulatory network*", but not a *sine qua non* condition) and fill the gaps to ensure that guidelines are meaningfully updated and properly anchored in clinical practice. For data collection, which should be aligned with the estimands, the case report forms (CRFs) and data collection standards should also be aligned to accommodate meaningful data collection, e.g. intercurrent events data [46].

Furthermore, the sponsors should update protocol and statistical analysis plan templates to integrate the estimands framework and enable its application according to ongoing initiatives [34,35]. Case studies and simulations are needed at each disease level or in clusters of diseases with common grounds. For alignment with the estimands and proportional improvement of clinical trial reporting and interpretation, the clinical study report (CSR) templates should be updated to integrate the estimands framework.

All healthcare stakeholders

Furthermore, every healthcare stakeholder should be concerned and contribute to the in-depth application of the estimands framework and extend its recommendations in drug development for more precise and better-defined treatment effects.

The COVID-19 pandemic has driven many changes to planned clinical trials, and the estimands framework was seen and used as a possible solution. However, there are no published trials yet that report in-depth the challenges encountered and how the estimands framework was used as a solution to address the possible impact of the pandemic (e.g. intercurrent events related to the COVID-19 pandemic). This is a missed opportunity for the scientific community. Timely scientific communication of value learnings from trials facing challenges due to pandemics or other disruptive events should be an objective for the benefit of other trials and, ultimately, for all patients.

Converging estimands framework with causal inference

Another point of attention should be the (apparently) disjunct worlds of regulatory statistics and causal inference [25,47]. Although the ICH E9(R1) addendum does not contain the word "causal", it does refer to estimands that are also causal. Even though there are randomised and non-randomised settings, the ICH E9(R1) addendum suggests a principal stratum strategy for addressing intercurrent events. To estimate such an estimand prespecified with a principal stratum strategy, causal inference methodology needs to be employed. This methodology is the link; it can be used for any setting, not conditional on the randomised/non-randomised settings, while the question is usually similar to "What is the treatment effect in those patients that can comply with the treatment regimen or that will be able to take the treatment without toxicity or that will not experience AEs". Few initiatives exist [21,22,48–50]; still, the medical and scientific community could better collaborate to bridge the different practices and extend the methodology armamentarium to solve problems and address challenges when needing to estimate treatment effects in studies to benefit patients' healthcare [48]. The estimands framework was developed for randomised clinical trials, but the methodology can be applied to non-randomised settings [51], where there is at least a control condition included. This extension of application could be co-led by researchers from both fields.

Concluding observations

Implementation of estimands can bring value in the entire process of drug development, from the design of clinical trials to communicating treatment effects to patients.

This thesis can help the implementation and application of the estimands framework by using an explicit structure of estimands constructs, helping to estimate formulated estimands, and understanding what treatment effects can correspond to formulated estimands. The proposed DGMs are a powerful tool for evaluating estimands and understanding possible ways to model the association between outcomes and intercurrent events. This can provide valuable insights into the way they impact the targeted estimands.

The estimands framework can impact all healthcare stakeholders at different levels. Its impact depends on the role and timing of their involvement in the life cycle of medicinal product development. For instance, it can impact patients and caretakers because they are ultimately the receiver of the clinical benefit and the best positioned to judge whether the estimands are meaningful for them.

The estimands framework was developed with a focus on efficacy, but the framework and adverse events have also crossed paths [52]. However, since the approval of new medicinal products depends on the benefit-risk assessment [53], one could wonder whether the estimands should be considered for the duality efficacy-safety in scientific evaluation and regulatory approval. Furthermore, there should be more clarity and a bridging between pre- and post-approval. For instance, should the registry-based studies, conducted following approval, target the same or different estimands as the randomised controlled trials conducted to support the regulatory approval? [25,54] Drug development could benefit from more clarity on this matter.

Estimands may not be easy to implement, as they require time and a thinking process, but they are necessary. Estimands require genuine engagement in discussion and interaction between different stakeholders and professions, from regulators to patients and from statisticians to clinicians. This is needed to ensure that targeted estimands are meaningful, of genuine interest, and the driver of evaluation, approval, prescription, and clinical practice. They demand a multidisciplinary approach and great attention to hidden and embedded details that need to be addressed thoroughly to (better) understand treatment effects in clinical trials. They may tend to make clinical trials more complicated at the beginning (planning and design stages) but make the subsequent steps smooth in the end, and the net beneficial impact on drug development should

be undisputed. Ideally, no researcher, regulator, statistician, patient, or trialist should have to reverse-engineer estimands in the near future. They should be available and explicitly described in study protocols, statistical analysis plans, clinical study reports and synopses, and public assessment reports.

References

1. Temple R, Ellenberg SS. Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments. Part 1: Ethical and Scientific Issues. Ann Intern Med. 2000 Sep 19;133(6):455–63.
2. Campbell RT, Willox GP, Jhund PS, Hawkins NM, Huang F, Petrie MC, et al. Reporting of Lost to Follow-Up and Treatment Discontinuation in Pharmacotherapy and Device Trials in Chronic Heart Failure: A Systematic Review. Circ Heart Fail [Internet]. 2016 May [cited 2021 Jun 23];9(5). Available from: https://www.ahajournals.org/doi/10.1161/CIRCHEARTFAILURE.115.002842
3. Permutt T. A taxonomy of estimands for regulatory clinical trials with discontinuations. Stat Med. 2016 Jul 30;35(17):2865–75.
4. Little R, Kang S. Intention-to-treat analysis with treatment discontinuation and missing data in clinical trials. Stat Med. 2015 Jul 20;34(16):2381–90.
5. Mehrotra DV, Hemmings RJ, Russek-Cohen E. Seeking harmony: estimands and sensitivity analyses for confirmatory clinical trials. Clin Trials. 2016 Aug 1;13(4):456–8.
6. ICH E9(R1) EWG. Final Concept Paper E9(R1): Addendum to Statistical Principles for Clinical Trials on Choosing Appropriate Estimands and Defining Sensitivity Analyses in Clinical Trials, Oct. 2014 [Internet]. 2014 [cited 2018 Jan 17]. Available from: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9__R1__Final_Concept_Paper_October_23_2014.pdf
7. Beckett RD, Loeser KC, Bowman KR, Towne TG. Intention-to-treat and transparency of related practices in randomized, controlled trials of anti-infectives. BMC Med Res Methodol. 2016 24;16(1):106.
8. Gravel J, Opatrny L, Shapiro S. The intention-to-treat approach in randomized controlled trials: are authors saying what they do and doing what they say? Clin Trials Lond Engl. 2007;4(4):350–6.
9. Day S. Editorial: Estimands in clinical trials – a challenge to intention to treat? Stat Med. 2017;36(1):3–4.
10. DeSantis SM, Li R, Zhang Y, Wang X, Vernon SW, Tilley BC, et al. Intent-to-treat analysis of cluster randomized trials when clusters report unidentifiable outcome proportions. Clin Trials. 2020 Dec 1;17(6):627–36.
11. Alshurafa M, Briel M, Akl EA, Haines T, Moayyedi P, Gentles SJ, et al. Inconsistent Definitions for Intention-To-Treat in Relation to Missing Outcome Data: Systematic Review of the Methods Literature. PLOS ONE. 2012 Nov 15;7(11):e49163.
12. Leuchs AK, Brandt A, Zinserling J, Benda N. Disentangling estimands and the intention-to-treat principle. Pharm Stat. 2017 Jan 1;16(1):12–9.
13. Oude Rengerink K, Mitroiu M, Teerenstra S, Pétavy F, Roes KCB. Rethinking the intention-to-treat principle: one size does not fit all. J Clin Epidemiol. 2020 Sep;125:198–200.
14. Keene ON, Ruberg S, Schacht A, Akacha M, Lawrance R, Berglind A, et al. What matters most? Different stakeholder perspectives on estimands for an invented case study in COPD. Pharm Stat. 2020;19(4):370–87.
15. Keene ON, Wright D, Phillips A, Wright M. Why ITT analysis is not always the answer for estimating treatment effects in clinical trials. Contemp Clin Trials. 2021 Jun 26;106494.
16. Callegari F, Akacha M, Quarg P, Pandhi S, von Raison F, Zuber E. Estimands in a Chronic Pain Trial: Challenges and Opportunities. Stat Biopharm Res. 2019 Jul 22;1–6.
17. Akacha M. Choosing Measures of Treatment Benefit: Estimands and Beyond. CHANCE. 2019 Oct 2;32(4):12–7.
18. Akacha M, Bretz F, Ruberg S. Estimands in clinical trials - broadening the perspective. Stat Med. 2017 Jan 15;36(1):5–19.
19. Akacha M, Kothny W. Estimands: A More Strategic Approach to Study Design and Analysis. Clin Pharmacol Ther. 2017 Dec 1;102(6):894–6.
20. Mehrotra DV. A note on the draft International Council for Harmonisation guidance on estimands and sensitivity analysis. Clin Trials. 2019 Apr 12;174077451984425.
21. Bornkamp B, Rufibach K, Lin J, Liu Y, Mehrotra DV, Roychoudhury S, et al. Principal stratum strategy: Potential role in drug development. Pharm Stat. 2021 Feb 23;pst.2104.
22. Magnusson BP, Schmidli H, Rouyrre N, Scharfstein DO. Bayesian inference for a principal stratum estimand to assess the treatment effect in a subgroup characterized by postrandomization event occurrence. Stat Med. 2019;38(23):4761–71.
23. Qu Y, Luo J, Ruberg SJ. Implementation of tripartite estimands using adherence causal estimators under the causal inference framework. Pharm Stat. 2021;20(1):55–67.

24. Qu Y, Shurzinske L, Sethuraman S. Defining estimands using a mix of strategies to handle intercurrent events in clinical trials. Pharm Stat [Internet]. [cited 2020 Oct 25];n/a(n/a). Available from: http://onlinelibrary.wiley.com/doi/abs/10.1002/pst.2078

25. ICH E9(R1) EWG. ICH E9(R1) Estimands and Sensitivity Analysis in Clinical Trials. STEP 4 TECHNICAL DOCUMENT. [Internet]. [cited 2019 Dec 4]. Available from: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf

26. Keene ON, Ruberg S, Schacht A, Akacha M, Lawrance R, Berglind A, et al. What matters most? Different stakeholder perspectives on estimands for an invented case study in COPD. Pharm Stat. 2020 Jan 9;

27. Akacha M, Bartels C, Bornkamp B, Bretz F, Coello N, Dumortier T, et al. Estimands—What they are and why they are important for pharmacometricians. CPT Pharmacomet Syst Pharmacol. 2021;10(4):279–82.

28. Kang M, Kendall MA, Ribaudo H, Tierney C, Zheng L, Smeaton L, et al. Incorporating estimands into clinical trial statistical analysis plans. Clin Trials. 2022 Mar 8;174077452210804.

29. EMA. Appendix 2 to the guideline on evaluation of anticancer medicinal products in man - use patient-reported outcome (PRO) measures oncology studies [Internet]. European Medicines Agency. 2018 [cited 2022 Jun 6]. Available from: ttps://www.ema.europa.eu/en/appendix-2-guideline-evaluation-anticancer-medicinal-products-man-use-patient-reported-outcome-pro

30. Lawrance R, Degtyarev E, Griffiths P, Trask P, Lau H, D'Alessio D, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? J Patient-Rep Outcomes. 2020 Dec;4(1):68.

31. EMA. Clinical efficacy and safety guidelines [Internet]. European Medicines Agency. 2018 [cited 2022 Apr 19]. Available from: https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-guidelines/clinical-efficacy-safety-guidelines

32. Commissioner O of the. Search for FDA Guidance Documents [Internet]. FDA. FDA; 2022 [cited 2022 May 14]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents

33. Real-world evidence framework feedback | What we do | About [Internet]. NICE. NICE; [cited 2022 Jun 6]. Available from: https://www.nice.org.uk/about/what-we-do/real-world-evidence-framework-feedback

34. ICH Official web site: ICH [Internet]. [cited 2019 Oct 14]. Available from: https://www.ich.org/page/multidisciplinary-guidelines

35. TransCelerate - Clinical Content & Reuse Assets - Clinical Studies [Internet]. TransCelerate. [cited 2022 May 15]. Available from: https://www.transceleratebiopharmainc.com/assets/clinical-content-reuse-solutions/

36. Consort - Welcome to the CONSORT Website [Internet]. [cited 2022 Feb 27]. Available from: http://www.consort-statement.org/

37. Dimairo M, Pallmann P, Wason J, Todd S, Jaki T, Julious SA, et al. The Adaptive designs CONSORT Extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design. BMJ. 2020 Jun 17;369:m115.

38. RoB 2: A revised Cochrane risk-of-bias tool for randomized trials | Cochrane Bias [Internet]. [cited 2022 Aug 1]. Available from: https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool -randomized-trials

39. EMA. Evaluation of anticancer medicinal products in man [Internet]. European Medicines Agency. 2018 [cited 2022 May 17]. Available from: https://www.ema.europa.eu/en/evaluation-anticancer-medicinal-products-man

40. EMA. Assessment templates and guidance [Internet]. European Medicines Agency. 2018 [cited 2022 May 16]. Available from: https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/assessment-templates-guidance

41. EMA. European public assessment reports: background and context [Internet]. European Medicines Agency. 2018 [cited 2022 May 16]. Available from: https://www.ema.europa.eu/en/medicines/what-we-publish-when/european-public-assessment-reports-background-context

42. EMA. Summary of product characteristics [Internet]. European Medicines Agency. [cited 2022 May 16]. Available from: https://www.ema.europa.eu/en/glossary/summary-product-characteristics

43. Home - ClinicalTrials.gov [Internet]. [cited 2017 Dec 10]. Available from: https://clinicaltrials.gov/

44. Oncology estimand working group [Internet]. [cited 2021 May 17]. Available from: http://www.oncoestimand.org

45. Neuroscience-Estimands European SIG [Internet]. PSI. [cited 2021 May 17]. Available from: https://www.psiweb.org/sigs-special-interest-groups/neuroscience-estimands

46. Implementation of Estimands (ICH E9(R1)) in Data Standards - WORKING GROUPS - PHUSE Advance Hub [Internet]. [cited 2022 Mar 29]. Available from: https://advance.phuse.global/display/WEL/Implementation+of+Estimands+%28ICH+E9+%28R1%29%29+in+Data+Standards

47. Murray EJ, Swanson SA, Hernán MA. Guidelines for estimating causal effects in pragmatic randomized trials. ArXiv191106030 Stat [Internet]. 2019 Nov 19 [cited 2022 Jan 24]; Available from: http://arxiv.org/abs/1911.06030

48. Lipkovich I, Ratitch B, Mallinckrodt CH. Causal Inference and Estimands in Clinical Trials. Stat Biopharm Res. 2020 Jan 2;12(1):54–67.

49. Michiels H, Sotto C, Vandebosch A, Vansteelandt S. A novel estimand to adjust for rescue treatment in randomized clinical trials. Stat Med. 2021 Feb 10;sim.8901.

50. Larsen KG, Josiassen MK. A New Principal Stratum Estimand Investigating the Treatment Effect in Patients Who Would Comply, If Treated With a Specific Treatment. Stat Biopharm Res. 2020 Jan 2;12(1):29–38.

51. Li H, Wang C, Chen W, Lu N, Song C, Tiwari R, et al. Estimands in observational studies: Some considerations beyond ICH E9(R1). Pharm Stat. 2022 Feb 6;pst.2196.

52. Unkel S, Amiri M, Benda N, Beyersmann J, Knoerzer D, Kupas K, et al. On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. Pharm Stat. 2019 Mar;18(2):166–83.

53. EMA. Benefit-risk methodology [Internet]. European Medicines Agency. 2018 [cited 2022 May 17]. Available from: https://www.ema.europa.eu/en/about-us/what-we-do/regulatory-science-research/benefit-risk-methodology

54. EMA. Guideline on registry-based studies [Internet]. European Medicines Agency. 2021 [cited 2022 Jun 6]. Available from: https://www.ema.europa.eu/en/guideline-registry-based-studies-0

APPENDICES

Samenvatting in het Nederlands

In dit proefschrift heb ik het (impliciete) gebruik van estimands in het pre-estimand tijdperk in klinische onderzoek onderzocht voor regulatoire evaluatie en goedkeuring, om te begrijpen welke estimands overeenkomen met veelgebruikte analyses van werkzaamheid. Daarnaast heb ik, om de kenmerken van schattingen beter te begrijpen, datagenererende modellen (DGM's) ontwikkeld die rekening houden met de associatie tussen klinische uitkomsten en tussentijdse gebeurtenissen.

In **Hoofdstuk 1** verwijs ik naar de eerste gerandomiseerde gecontroleerde studie die gerapporteerd is in de wetenschappelijke literatuur en leg ik het principe van randomisatie uit. Ik vat de ontwikkeling van het ICH E9(R1) estimands raamwerk en het estimand concept samen, en ik presenteer de doelstellingen en de hoofdlijnen van dit proefschrift.

In **Hoofdstuk 2** pleit ik in brede zin voor de noodzaak om behandeleffecten in klinisch onderzoek duidelijker te definiëren. Ik beargumenteer dat Intention-To-Treat en Per Protocol analyses niet alle relevante wetenschappelijke vragen (zonder vertekening) beantwoorden. Het *estimands framework* is van belang in de gezondheidszorg omdat het aanmoedigt om behandeleffect en beter te definiëren dan gebruikelijk op basis van Intention-To-Treat en Per Protocol analyses. Ik pleit voor een verdere discussie in de wetenschappelijke gemeenschap om te blijven leren en ervoor te zorgen dat de implementatie van het estimands raamwerk het ontwerpen en uitvoeren van klinische trials verbetert.

In **Hoofdstuk 3** evalueer ik bestaande methoden om effecten te schatten bij de ontwikkeling en evaluatie van geneesmiddelen, bekeken vanuit het estimands perspectief. De vijf attributen (behandeling, populatie, variabele, samenvatting op populatieniveau en strategieën voor tussentijdse gebeurtenissen die nog niet op het niveau van de genoemde attributen zijn vastgelegd) waren aanwezig in opzet en analyse van klinische onderzoek vóór de introductie van het estimand raamwerk, maar niet op een expliciete manier. Ik heb de beoogde estimand afgeleid uit de bestaande EMA-richtlijnen voor klinische werkzaamheid, klinische studierapporten en statistische analyseplannen die door de sponsor zijn ingediend ter ondersteuning van de aanvraag voor een handelsvergunning, aangevuld met de informatie uit de lijst met vragen (als onderdeel van de toelatingsprocedure) van de regelgevers. Nieuw aan het estimands raamwerk is de nauwkeurige beschrijving van behandelingseffecten en het expliciete gebruik van strategieën voor tussentijdse gebeurtenissen die zijn geïntegreerd in de primaire uitkomst of schatting.

Het extraheren en afleiden van de relevante tussentijdse gebeurtenissen was moeilijk omdat informatie met betrekking tot de estimands attributen verspreid was over de

klinische studierapporten en statistische analyseplannen, en verschilde tussen klinische studierapporten en dossiers voor geneesmiddelenontwikkeling. De strategieën voor het omgaan met tussentijdse gebeurtenissen kon niet uit alle regulatoire vragen worden afgeleid. De "*treatment policy*" strategie werd het vaakst geadviseerd in de EMA-richtlijnen voor klinische werkzaamheid, terwijl de "hypothetische strategie" het vaakst werd toegepast in de sponsordocumentatie, hoewel deze in geen enkele ziekte specifieke richtlijn werd geadviseerd. De resultaten geven aan dat het regulatoire doel van schatting en wat feitelijk wordt geschat niet volledig overeenstemmen, voornamelijk als gevolg van beperkingen bij het verzamelen van uitkomstgegevens om een "*treatment policy*" strategie voor alle tussentijdse gebeurtenissenmogelijk te maken. Het onderzoek voor dit hoofdstuk toonde aan dat het nodig was om te beginnen met een nauwkeurige beschrijving van de beoogde behandelingseffecten en om de onderzoeksopzet en -analyse op daarop af te stemmen. Hiertoe is verdere training van betrokken professionals en opname van estimands in richtlijnen voor klinische werkzaamheid nodig, op basis van multidisciplinaire discussies, om een zinvoller schattingsdoel voor verschillende belanghebbenden nauwkeuriger te definiëren. Het toonde ook aan dat wanneer uitkomsten en registraties (deels) ontbreken, het gebruik van "*treatment policy*" in strikte zin niet mogelijk is. Als dit wel het beoogde doel is, moet dit expliciet in opzet en uitvoering worden meegenomen.

In hoofdstuk 3 heb ik vanuit een estimands perspectief gekeken naar methoden om effecten te schatten bij de ontwikkeling en evaluatie van geneesmiddelen, zoals ze vastliggen in EMA-richtlijnen voor klinische werkzaamheid, protocollen van de sponsor, plannen voor statistische analyse en klinische studierapporten, en regulatoire vragen tijdens wetenschappelijke evaluatie. In **Hoofdstuk 4** heb ik vanuit een estimands perspectief gekeken naar veelgebruikte analysemethoden met behulp van concrete data van patiënten uit klinische trials. Ik heb zes gerandomiseerde gecontroleerde onderzoeken opnieuw geanalyseerd die een nieuwe behandeling tegen depressie bij volwassenen evalueerden, om te begrijpen welke estimands overeenkomen met veelgebruikte werkzaamheidsanalyses en om het verschil in de grootte van de schatting van de verschillende gebruikte analysemethoden te beoordelen. Ik heb de volgende analysemethoden geselecteerd: Analyse van covariantie (ANCOVA) op data van proefpersonen zonder missende waarden, ANCOVA na *last observation carried forward* (LOCF, laatst gemeten waarde wordt voortgezet) imputatie of multipele imputatie, mixed models voor herhaalde metingen (MMRM), MMRM na LOCF-imputatie of MMRM na *jump-to-reference* (sprong-naar-referentie) imputatie, en "*pattern-mixture mixed models*". Ik heb ook een "*principal stratum*" analyse opgenomen van een stratum (van belang) van de onderzoekspopulatie die niet zou stoppen vanwege bijwerkingen of gebrek aan werkzaamheid. Ik vertaalde elke analyse naar de beoogde estimands en de bijbe-

horende klinische vragen. Ik ontdekte dat dezelfde analysemethode gekoppeld kon worden aan meer dan één estimand, en dezelfde schatting gekoppeld kon worden aan meer dan één analyse. De belangrijkste verschillen tussen estimands betreffen de keuze van de strategie voor elke tussentijdse gebeurtenis.

Niet alle zes estimands hadden een klinisch betekenisvolle interpretatie. Dit bevestigt de noodzaak om estimands te formuleren bij de opzet van een klinisch onderzoek in een multidisciplinair team, waarbij vooral clinici en statistici betrokken zijn. Imputaties (voor missende waarden) worden vaak gebruikt in combinatie met analysemethoden zonder rekening te houden met de afstemming van de geïmputeerde waarden op het onderzoeksdoel. Een andere belangrijke constatering is dat er geen onderscheid wordt gemaakt tussen de redenen die hebben geleid tot ontbrekende gegevens, zoals een gebrek aan werkzaamheid of een bijwerking. Beide worden gewoonlijk op dezelfde manier geïmputeerd, terwijl voor een goede klinische interpretatie verschil tussen beide oorzaken gemaakt zou moeten worden bij het bepalen van een imputatiestrategie. Slechts enkele analyses zijn op dezelfde estimand gericht; daarom kunnen er per definitie maar weinig als elkaars sensitiviteitsanalyses worden gebruikt. Sommige analyses die als sensitiviteitsanalyses werden gebruikt en gerapporteerd, zijn volgens ICH E9(R1) geen echte sensitiviteitsanalyses, aangezien ze gericht zijn op een andere estimand in plaats van de gevoeligheid van aannames voor de schatting van dezelfde estimand te testen. Sommige analyses waren gericht op dezelfde estimand, maar het waren geen sensitiviteitsanalyses van elkaar omdat ze dezelfde onderliggende aannames hadden. ANCOVA die MAR meervoudige imputatie volgt en MMRM zijn bijvoorbeeld gebaseerd op dezelfde (statistische) aannames.

Onze bevindingen benadrukken dat de estimands vooraf moeten worden gespecificeerd. Het reconstrueren van de beoogde estimand op basis van de toegepaste analyse methoden p, is niet eenduidig mogelijk, omdat er geen 1-op-1 overeenkomst is tussen gangbare analysemethoden en estimands.

In **Hoofdstuk 5** had ik als doel data genererende modellen (DGMs) te ontwikkelen en te evalueren om gezamenlijk uitkomsten en tussentijdse gebeurtenissen voor gerandomiseerde klinische onderzoek te kunnen simuleren. Ik stelde vier DGM's voor, voor het genereren van de gezamenlijke verdeling van longitudinale continue klinische uitkomsten entussentijdse gebeurtenissen onder het scenario waarin ze waarneembaar zijn: een selectiemodel, een "pattern mixture" model, een "*shared-parameter*" model, en een "*joint model*" van longitudinaal waargenomen klinische uitkomsten en een overlevingsduur model voor tussentijdse gebeurtenissen. Als casestudy gebruikte ik een kortdurende depressiestudie met herhaalde metingen van continue uitkomsten en

twee soorten gebeurteissen (gebrek aan werkzaamheid en bijwerkingen) en vergeleek de vier voorgestelde DGM's. Ik stelde mogelijke manieren voor om deze DGM's te implementeren. Ik ontdekte dat alle vier DGM's een *target trial* (d.w.z. de studie die je zou willen uitvoeren) kunnen nabootsen in termen van beoogde patronen van tussentijdse gebeurtenissen en uitkomst patronen in de tijd met verschillende mate van precisie. Eerder beschikbare klinische onderzoeksgegevens waarop modellen en parameters kunnen worden gebaseerd, kunnen worden gebruikt om realistische *target trials* te bereiken.

De vier voorgestelde DGM's kunnen een breed scala aan scenario's simuleren die de evaluatie van verschillende estimands mogelijk maakt en een diepgaand onderzoek van hun eigenschappen mogelijk maakt. Vervolgens kunnen simulatiestudies met behulp van deze DGM's het fundamentele begrip van estimands vergroten en kunnen ze informatie verschaffen over planning, ontwerp, uitvoering en analyse van gerandomiseerde klinische onderzoeken, evenals interpretatie van de resultaten die door onderzoeken worden gegenereerd.

Implicatie voor de praktijk

Patiënten en andere zorgprofessionals

Ik breng onder de aandacht dat er een nieuwe wetenschappelijke richtlijn is om behandeleffecten in klinische onderzoek beter te beschrijven en in te schatten. Ik pleit er ook voor dat de momenteel veel gebruikte en gerapporteerde behandeleffecten niet de enige behandeleffecten zijn die van belang zijn.

Het estimands raamwerk opent nieuwe mogelijkheden in de ontwikkeling van geneesmiddelen en kan worden gebruikt als een kans voor patiënten om hun eigen cruciale rol in klinische onderzoek vorm te geven, om meer betrokken te zijn bij de besluitvorming in alle stadia van de ontwikkeling van nieuwe behandelingen. De mening van patiënten kan beter worden geïntegreerd in onderzoeken die worden uitgevoerd naar de ziekten waaraan ze lijden of in onderzoeken waaraan ze deelnemen. Er zijn meer estimands, en sommige kunnen interessanter zijn voor patiënten, terwijl ze niet van primair belang zijn voor andere belanghebbenden, zoals regelgevers of instituten die bepalen over de vergoedinging van geneesmiddelen. Het is onduidelijk wat de waarde is van Intention-To-Treat of Per Protocol analyses voor patiënten. Er zijn andere behandelingseffecten die mogelijk interessanter zijn voor patiënten.

Patiënten bevinden zich in een positie om klinisch onderzoek erichting te geven naar het schatten van behandelingseffecten die zinvoller zijn, omdat zij het beste weten wat werkt en wat voor hen belangrijk is. Via de principale stratumstrategie zouden patiënten bijvoorbeeld de klinische vraagstelling en de bijbehorende estimands veel duidelijker kunnen sturen omdat er meer opties worden gecreëerd. Zo kunnen ze, prominenter en meer vanaf het beginbij dit debat betrokken worden en de onderzoeksdoelstelling mede bepalen.

Er ontstaan ook kansen voor bijvoorbeeld *Patient-Reported Outcomes* (PRO), een gebied dat baat zou kunnen hebben bij integratie met het E9(R1) estimands raamwerk.

Biostatistici en clinici

Estimands zijn in het kader van doelstelling van klinisch onderzoek een grote verantwoordelijkheid die moet worden gedeeld tussen biostatistici en clinici die betrokken zijn vanaf de beginfase van het onderzoek, van uitvoering en analyse tot interpretatie van resultaten. De in dit proefschrift voorgestelde DGM's zijn een nieuwe manier om schattingen diepgaand te onderzoeken met simulatiestudies. Ze bieden modelering van het verband tussen uitkomsten en tussentijdse gebeurtenissen, en kunnen worden gebruikt om de eigenschappen van verschillende mogelijke estimands van belang voor verschillende belanghebbenden te evalueren. Ze kunnen worden geïnformeerd door andere onderzoeken en door deskundige kennis van klinische onderwerpen. Multidisciplinaire samenwerking is van belang om de onderzoeksdoelstellingen, gegevensverzameling, analyses, interpretatie en communicatie van onderzoeksresultaten op elkaar af te stemmen, waarbij in elke fase gebruik wordt gemaakt van het estimands raamwerk.

Geneesmiddelenregulatoren en geneesmiddelenvergoedingsinstanties, als organisaties

Ik pleit voor het estimands raamwerk en ik heb constructieve kritiek op Intention-To-Treat en Per Protocol analyses. Hoewel ze nuttig waren vóór het estimands raamwerk is vastgesteld werd het duidelijk dat ze niet altijd geschikt zijn of niet de enige behandelingseffecten zijn die van belang zijn in klinische onderzoek en voor alle belanghebbenden.

Enig impliciet estimands denken is al aanwezig in richtlijnen voor geneesmiddelenontwikkeling. Ik voorzie een algemene en volledige herziening van richtlijnen voor klinische werkzaamheid en een verbeterde methodologie, met expliciet gebruik van het estimands addendum, in het kader van klinische onderzoeken ter ondersteuning van aanvragen voor handelsvergunningen. De toepassing en integratie van het estimands raamwerk bij de ontwikkeling van geneesmiddelen bevindt zich echter nog

in de beginfase. Het is misschien niet gemakkelijk en het zal waarschijnlijk ook niet snel gebeuren. Toch is het zeker nodig om te zorgen voor afstemming met het ICH E9(R1) addendum om op estimands gebaseerde geneesmiddelenontwikkeling en regelgevende evaluatie te vergemakkelijken.

Het is niet ongebruikelijk dat verschillende regelgevende instanties wetenschappelijk advies of protocolondersteuning kunnen geven met verschillende aanbevelingen aan sponsors die van plan zijn een bepaald onderzoek uit te voeren. Door de nauwkeurigheid van gerichte behandeleffecten te verbeteren, leidt het estimands raamwerk mogelijk tot meer geharmoniseerde adviezen. Het zou er in ieder geval toe moeten leiden dat er meer duidelijkheid komt voor afwijkende of convergerende aanbevelingen en dat de verschillende gemaakte keuzes worden verantwoord en dat er aanbevelingen worden gedaan om tot een definitief advies te komen.

Ik voorzie dat de verstrekking van wetenschappelijk advies en protocolondersteuning veel beter en duidelijker zal worden, op voorwaarde dat de verzoeken om advies gedetailleerde, door estimand definities ondersteunde en DGM's-geïnformeerde vragen bevatten. Mogelijk zorgt dit voor consistentie tussen de estimands die door de regelgevende instanties worden geadviseerd en de estimands die door sponsors worden gebruikt. De voorgestelde DGM's kunnen ook worden gebruikt om meerdere estimands die van belang zijn in dezelfde studie(s) te evalueren. Bijvoorbeeld sommige estimands die moeten worden gebruikt voor evaluatie van de werkzaamheid en goedkeuring door de regelgevende instanties, andere estimands kunnen worden gebruikt voor prijsstelling, en een of meer estimands die moeten worden gebruikt voor de therapeutische beslissingen van voorschrijvende artsen. Er is veel onderzoek gepubliceerd naar estimands die in klinische onderzoek kunnen worden gebruikt, maar de nadruk ligt op primair aantonen van werkzaamheid; er is minder aandacht voor deze bredere dimensie van estimands. Verder is er geen formeel standpunt ingenomen door HTA-organisaties over de toepassing en mogelijke meerwaarde van het estimands raamwerk. Dit is onbekend terrein dat in kaart zou kunnen worden gebracht met behulp van de voorgestelde DGM's en casestudies.

De informatie met betrekking tot estimand attributen is verspreid over de protocollen, statistische analyseplannen en klinische onderzoeksrapporten. Ons onderzoek kan bijdragen om de lopende initiatieven te stimuleren om het estimands raamwerk te integreren in relevante templates (plannen voor statistische analyse, protocollen, samenvattingen, klinische onderzoeksrapporten). Deze reeks van infrastructurele veranderingen om schattingskaders te integreren in alle dimensies van geneesmiddelenontwikkeling is al gaande. Gesteund door de noodzaak vanuit regelgeving (weten-

schappelijke evaluatie en wettelijke goedkeuring), bereikt het waarschijnlijk snel regelgevende instanties en de industrie. Het bereikt mogelijk niet tegelijkertijd de academische wereld. Er zijn duidelijk uitdagingen, maar die kunnen worden overwonnen door een multidisciplinaire en *multistakeholder* benadering en samenwerking.

Rezumat în română

În această teză, am investigat utilizarea (implicită) a estimanzilor în era pre-estimand în studiile clinice pentru evaluarea și aprobarea reglementară, pentru a înțelege ce estimanzi pot corespunde analizelor comune pentru eficacitate. În plus, pentru a înțelege fundamental caracteristicile estimanzilor, am dezvoltat modele generatoare de date (DGM) care iau în considerare asocierea dintre valorile parametrilor clinici și evenimentele intercurente.

În **Capitolul 1**, fac referință la primul studiu controlat randomizat raportat în literatura științifică și explic principiul randomizării. Sumarizez dezvoltarea cadrului ICH E9(R1) pentru estimanzi și conceptul de estimand și prezint obiectivele și schița acestei teze.

În **Capitolul 2**, promovez la scară largă necesitatea de a defini mai clar efectele tratamentelor în cercetarea clinică. Critic în mod constructiv faptul că analizele *Intention-To-Treat* și *Per Protocol* nu răspund la toate întrebările științifice relevante (fără bias) și aduc în atenția părților interesate din domeniul sănătății faptul că acest cadru al estimanzilor încurajează vizarea și definirea efectelor tratamentelor dincolo de analizele *Intention-To-Treat* și *Per Protocol*. Susțin o discuție continuă în comunitatea științifică pentru a continua procesul de învățare și pentru a asigura că implementarea cadrului estimanzilor îmbunătățește practicile din studiile clinice.

În **Capitolul 3,** am evaluat practicile anterioare de estimare a efectelor tratamentelor în dezvoltarea și evaluarea medicamentelor, privite din perspectiva estimanzilor. Cele cinci atribute (tratament, populație, variabilă, rezumat la nivel de populație și strategii pentru alte evenimente intercurente necaptate încă la nivelul atributelor menționate) au fost prezente în studiile clinice înainte de introducerea cadrului estimanzilor, dar nu într-un mod explicit. Am aplicat o inginerie inversă asupra estimanzilor din ghidurile de eficacitate clinică EMA, rapoartelor de studii clinice și planurilor de analiză statistică prezentate de sponsor pentru a sprijini cererea de autorizare de punere pe piață și din lista de întrebări a autorităților de reglementare din timpul evaluării dosarului (întrebările reglementare). Noutatea pe care o introduce cadrul estimanzilor este descrierea precisă a efectelor tratamentelor și utilizarea explicită a strategiilor pentru evenimentele intercurente integrate în rezultatul sau estimarea primară.

Extragerea și derivarea evenimentelor intercurente a fost dificilă deoarece informațiile relevante referitoare la atributele estimanzilor au fost dispersate între rapoartele studiilor clinice și planurile de analiză statistică, în mod eterogen între rapoartele studiilor clinice și între planurile dezvoltării medicamentelor. Strategiile pentru evenimente intercurente nu au putut fi derivate din toate întrebările reglementare. Strategia *treatment policy* a fost cel mai adesea recomandată în ghidurile EMA privind eficacitatea clinică, în timp ce strategia

hypothetical a fost strategia principală aplicată în documentația sponsorului, în ciuda faptului că nu a fost recomandată în niciun ghid de efficacitate clinică. Rezultatele indică faptul că ținta reglementară a estimării și ceea ce este estimat de fapt nu sunt pe deplin concordante, în primul rând din cauza limitărilor în colectarea datelor parametrilor clinici pentru a permite o strategie *treatment policy* pentru toate evenimentele intercurente.

Cercetarea pentru acest capitol a arătat necesitatea de a începe cu o descriere precisă a efectelor tratamentului vizat și de a alinia design-ul și analiza studiului. În acest scop, sunt necesare diseminarea, instruirea și implementarea în continuare a estimanzilor în ghidurile de eficacitate clinică bazate pe discuții multidisciplinare pentru a defini mai precis o țintă relevantă de estimare pentru diferite părți interesate. De asemenea, a arătat că atunci când datele clinice lipsesc, deși *treatment policy* poate fi țința, utilizarea strategiei *treatment policy* nu este posibilă în sensul strict și acest lucru trebuie abordat în mod explicit.

În Capitolul 3, am analizat din perspectiva estimanzilor practicile de estimare în dezvoltarea și evaluarea medicamentelor în formă scrisă (ghidurile EMA privind eficacitatea clinică, protocoalele sponsorului, planuri de analiză statistică și rapoarte de studii clinice și întrebări de reglementare în timpul evaluării științifice). În **Capitolul 4**, am investigat din perspectiva estimanzilor metodele de analiză comune folosind date ale pacienților din studii concrete. Am reanalizat șase studii randomizate controlate care evaluează un nou tratament antidepresiv la adulți pentru a înțelege ce estimanzi corespund analizelor comune pentru eficacitate și pentru a evalua diferența în dimensiunea efectului clinic al tratamentelor estimate de diferitele metode de analiză utilizate. Am selectat următoarele metode de analiză: *ANCOVA* pe cazuri complete (*complete cases*), *ANCOVA* după imputarea *LOCF* sau după imputare multiplă (*multiple imputation*), *MMRM*, *MMRM* după imputarea *LOCF* sau *MMRM* după imputarea *jump-to-reference* și *PMMM*. Am inclus, de asemenea, o analiză de tip *principal stratum* pe un strat (de interes) al populației de studiu care nu ar întrerupe tratamentul din cauza evenimentelor adverse sau a lipsei de eficacitate. Am translatat fiecare analiză în estimanzii vizați și întrebările clinice corespunzătoare.

Am descoperit că aceeași metodă de analiză ar putea fi corelată cu mai mulți estimanzi și același estimand ar putea fi asociat cu mai multe analize. Diferențele majore între cei șase estimanzi constă în alegerea strategiei pentru fiecare eveniment intercurent.

Nu toți cei șase estimanzi au avut o interpretare semnificativă din punct de vedere clinic. Acest lucru sugerează necesitatea de a formula estimanzi în studiile clinice într-o echipă multidisciplinară, implicând în special clinicieni și statisticieni. Imputările sunt adesea folosite împreună cu metodele de analiză fără a se lua în considerare alinierea valorilor rezultatelor clinice imputate cu obiectivul studiului. O altă constatare importantă este că

nu se face distincție între motivele care au condus la lipsa datelor clinice, cum ar fi lipsa de eficacitate sau un eveniment advers. Ambele sunt de obicei imputate în același mod, în timp ce în ceea ce privește interpretarea clinică, cele două ar trebui considerate diferit în determinarea oricărei strategii de imputare.

Doar câteva analize ar viza același estimand; prin urmare, prin definiție, puține ar putea fi utilizate ca *sensitivity analyses*. Unele analize utilizate și comunicate ca *sensitivity analyses* nu au fost cu adevărat *sensitivity analyses* conform ICH E9(R1), deoarece vizează un estimand diferit, mai degrabă decât testarea sensibilității ipotezelor pentru estimarea aceluiași estimand. Unele analize au vizat același estimand, dar nu au fost *sensitivity analyses* reciproc, deoarece aveau aceleași ipoteze de bază. De exemplu, *ANCOVA* după imputarea multiplă *MAR* și *MMRM* se bazează pe aceleași ipoteze.

Descoperirile noastre subliniază faptul că estimanzii ar trebui să fie prespecificați, deoarece atunci când estimanzii sunt reconstruiți, nu există o mapare 1-la-1 între metodele de analiză comune și estimanzi.

În **Capitolul 5,** mi-am propus să dezvolt și să evaluez modele generatoare de date clinice pentru a simula în conjuncție date clinice și evenimentele intercurente pentru studiile clinice randomizate. Am propus patru modele generatoare de date clinice pentru distribuția comună a datelor clinice longitudinale continue și a evenimentelor intercurente în scenariul în care acestea sunt observabile (adică nu au avut loc evenimente terminale în timpul perioadei de observație): un *selection model*, un *pattern-mixture mixed model*, un *shared-parameter model*, un *joint model* al datelor clinice observate longitudinal și un *survival model* pentru evenimente intercurente. Am folosit un studiu de caz într-un studiu clinic pe termen scurt pentru depresie cu măsurători repetate ale datelor clinice continue și două tipuri de evenimente intercurente (lipsa eficacității și evenimentul advers) și am comparat cele patru modele generatoare de date clinice propuse. Am propus modalități posibile de implementare a acestor modele generatoare de date clinice. Am descoperit că toate cele patru modele generatoare de date clinice pot imita un studiu țintă în ceea ce privește modelele preconizate de evenimente intercurente și traiectorii ale datelor clinice cu diferite grade de precizie. Datele disponibile din studiile clinice deja conduse pe care se pot baza modelele și parametrii pot fi utilizate pentru a simula studii țintă realiste.

Cele patru modele generatoare de date clinice propuse pot simula o gamă largă de scenarii care pot permite evaluarea diferiților estimanzi și permit investigarea aprofundată a proprietăților acestora. În consecință, studiile de simulare care utilizează aceste modele generatoare de date clinice pot dezvolta înțelegerea fundamentală a

estimanzilor și pot informa planificarea, design-ul, desfășurarea și analiza studiilor clinice randomizate, precum și interpretarea rezultatelor generate de aceste studii.

Implicații pentru practică

Pacienți și alți profesioniști din domeniul sănătății

Aduc în atenția comunității din domeniul sănătății că există un nou ghid științific pentru a descrie și estima mai bine efectele tratamentelor în studiile clinice. De asemenea, susțin că efectele tratamentelor des utilizate și raportate în prezent nu sunt singurele efecte de interes.

Cadrul estimanzilor deschide noi oportunități în dezvoltarea medicamentelor și poate fi folosit ca o oportunitate pentru pacienți de a-și modela propriul rol critic în studiile clinice, pentru a fi mai implicați în luarea deciziilor în toate etapele dezvoltării unui nou tratament. Opiniile pacienților pot fi mai bine integrate în studiile efectuate pentru bolile de care sunt afectați sau în studiile la care participă. Există mai mulți estimanzi, iar unii pot fi de interes pentru pacienți, fără a fi de interes principal pentru alte părți interesate, cum ar fi autoritățile de reglementare sau de rambursare a medicamentelor. Nu este clar care este valoarea analizelor *Intention-To-Treat* sau *Per Protocol* pentru pacienți; posibil, există și alte efecte ale tratamentelor de mai mare interes pentru pacienți. Acest lucru nu s-ar traduce neapărat prin exclusivitatea reciprocă a diferiților și posibil multiplilor estimanzi de interes, pentru diferitele părți interesate, în studiile clinice.

Pacienții sunt în măsură să conducă studiile pe căi de estimare a efectelor tratamentelor care sunt mai semnificative, deoarece ar ști cel mai bine ce funcționează și ce contează pentru ei. De exemplu, prin strategia *principal stratum*, pacienții ar putea ghida întrebările clinice și formularea estimanzilor corespunzători mult mai dominant, deoarece există mai multe opțiuni create și puse la dispoziție și pentru că ar putea fi implicați în această dezbatere mai proeminent și de la început, de la formularea obiectivul studiului si planificarea acestuia.

Sunt deschise și alte oportunități adiacente, de exemplu pentru rezultatele raportate de pacienți (PRO), care este o zonă ce ar putea beneficia de integrarea cu cadrul E9(R1) al estimanzilor.

Biostatisticieni și clinicieni

Estimanzii reprezintă o mare responsabilitate care trebuie împărțită între biostatisticienii și clinicienii implicați încă din etapele inițiale ale studiilor, de la efectuarea și analiza datelor clinice până la interpretarea rezultatelor. Modelele generatoare de date clinice propuse în această teză sunt o modalitate nouă de investigare aprofundată a estimanzilor prin intermediul studiilor de simulare. Acestea oferă control asupra asocierii dintre datele clinice și evenimentele intercurente și pot fi utilizate pentru a evalua proprietățile diferiților estimanzi de interes pentru diferitele părți interesate. Acești estimanzi pot fi informați prin alte studii și prin cunoștințele experților clinicieni. Trebuie asigurată colaborarea și interacțiunile multidisciplinare pentru a alinia obiectivele studiului, colectarea datelor clinice, analizele, interpretarea și comunicarea rezultatelor studiului, utilizând cadrul estimanzilor în fiecare etapă.

Autoritățile de reglementare și de rambursare a medicamentelor, ca organizații

Susțin cadrul estimanzilor și critic în mod constructiv analizele *Intention-To-Treat* și *Per Protocol*. Deși au fost utile înainte de cadrul estimanzilor, a devenit evident că fie nu sunt potrivite, fie nu sunt singurele efecte ale tratamentelor de interes în studiile clinice și pentru toate părțile interesate.

Deși unii estimanzi impliciți sunt deja prezenți în ghidurile de reglementare, am în vedere o reînnoire generală și completă a ghidurilor de eficacitate clinică și o metodologie îmbunătățită, cu utilizarea explicită a addendumului estimărilor, pentru studiile clinice care susțin cererile de autorizare de punere pe piață. Cu toate acestea, aplicarea și integrarea cadrului estimanzilor în dezvoltarea medicamentelor sunt încă în faza de început. S-ar putea să nu fie ușor și nici nu este probabil să se întâmple repede. Cu toate acestea, cu siguranță este necesar să se asigure alinierea cu addendumul ICH E9(R1) pentru a facilita dezvoltarea medicamentelor bazată pe estimanzi și evaluare reglementară.

Nu este neobișnuit ca diferite agenții de reglementare să ofere consiliere științifică sau asistență pentru protocol cu recomandări diferite sponsorilor care intenționează să efectueze un anumit studiu. Poate că prin îmbunătățirea preciziei efectelor tratamentelor vizate, cadrul estimanzilor poate să nu conducă neapărat la sfaturi științifice mai mult sau mai puțin armonizate. Totuși, ar trebui să conducă la mai multă claritate pentru recomandările divergente sau convergente și pentru justificarea diferitelor alegeri făcute, și ar trebui furnizate recomandări pentru a ajunge la un sfat final.

Mă gândesc că furnizarea de consiliere științifică și asistență pentru protocol să devină mult mai îmbunătățită și mai clară, cu condiția ca cererile de consiliere reglementară să conțină întrebări detaliate, susținute de estimanzi și informate de modele generatoare de date. Acest lucru ar putea asigura coerența între estimanzii recomandați de agențiile de reglementare și estimanzii utilizați de sponsori. Modele generatoare de date propuse pot fi, de asemenea, utilizate pentru a evalua estimanzi multipli de interes pentru aceleași studii. De exemplu, unii estimanzi care urmează să fie utilizați pentru evaluarea eficacității și aprobarea reglementară, alți estimanzi să fie utilizați pentru rambursare și stabilire a prețurilor și unul sau mai mulți estimanzi care urmează să fie utilizați pentru deciziile terapeutice ale medicilor care prescriu medicamentele. Există cercetări extensive publicate pentru estimanzi care urmează să fie utilizați în studiile clinice, dar accentul este pus pe eficacitate; există încă atenție mai puțină sau limitată asupra acestei dimensiuni a posibilei utilizări și valorii adăugate a estimanzilor. În plus, nu există nicio poziție formală adoptată de autoritățile de rambursare sau de organizațiile HTA cu privire la aplicarea și posibila valoare adăugată a cadrului estimanzilor. Acesta este un teritoriu neexplorat care ar putea fi mapat cu ajutorul modelelor generatoare de date și studiilor de caz propuse.

În plus, am descoperit că informațiile referitoare la atributele estimanzilor sunt dispersate în protocoale, planuri de analiză statistică și rapoarte de studii clinice. Cercetarea noastră poate contribui la baza actuală de cunoștințe ca nivel suplimentar pentru a conduce inițiativele curente în curs de integrare a cadrului estimanzilor în documentele (tipizate) importante și relevante (planuri de analiză statistică, protocoale, rezumate, rapoarte de studii clinice). Acest val de schimbări infrastructurale pentru integrarea cadrului estimanzilor în toate dimensiunile dezvoltării medicamentelor este în curs. Cu toate acestea, posibil din cauza imperativului contextului (evaluare științifică și aprobare reglementară), este probabil să ajungă rapid la agențiile de reglementare și la industria farmaceutică, dar este posibil să nu ajungă în mediul academic în același timp. Există unele provocări, dar ele ar putea fi depășite printr-o abordare și colaborare multidisciplinară și cu mai multe părți interesate. Acest lucru se întâmplă, dar nu în măsura potențialului existent.

Există încă un decalaj între pre-aprobare și post-aprobare, fiecare etapă cu provocări și oportunități diferite, pentru care comunitatea științifică trebuie să se angajeze în învățarea colaborativă.

Dankwoord

Prof. dr. K.C.B. Roes, dear Kit, I remember vividly your questions about missing data, ITT and PP analyses during our interview, and even more vividly, my imperfect and ambitious attempts to answer them. I am overjoyed that I had the opportunity for many more attempts at answering your initial questions and many, many more others throughout my PhD project. Any researcher deserves a PhD Promotor like you, but only a few receive this blessing. I am extremely fortunate and would like to express my high gratitude for your supervision and guidance during my PhD project and beyond. Thanks to your insights, thoughtful considerations, time, attention, and openness without hesitation, support, constructive critique, and the leadership you always showed, I have always learned more and more. You were like an angel investor who showed me patience and offered tireless support to help me develop into a better statistics and methodology researcher, and also a better person. I hope my work and the outcomes of my research and effort help the world at least a little because you helped me a lot. Thank you for the leap of faith and trust placed in me and my potential. I have learned flexibility, resilience, research rigour, trial methodology, visualisation, modeling, statistics and statistics history, self-confidence, and many more aspects that I tried applying in my research or implementing in my persona as trained and developing traits. Learning from you was a well of knowledge that still quenches my thirst for clinical trial methodology. Thank you!

Dr. K. Oude Rengerink, dear Katrien, we were supposed to meet the Asterix team in Barcelona, but that day had other plans. I found that time at Schiphol airport, from one gate to another, waiting room to another, and one terminal to another, was greatly valuable because it helped us establish rapport and build the foundations of open and fruitful collaboration. Thank you for your supervision and mentorship, thoughtful consideration, teaching me that less is more, especially in writing, and for helping my thoughts and arguments in manuscripts. Thank you for the walks in the botanical garden and along the Krommerijn, for the Epi MSc testimonial, for your emotional support and input, from the beginning to the very end, and for all the time you invested in elevating my skills. You have always been a reliable pillar throughout the multiple projects and PhD phases. I am forever grateful for everything you have done for me and my development!

Dr. S. Teerenstra, dear Steven, our first contact was at EMA in early 2016 when organising the first estimand workshop. Little did I know then that I would be so lucky to work with you again. I have never had the chance to benefit from more patience and kindness from a single person. I want to express my appreciation and gratitude for your effort to explain to me so many methodological matters, simple or complicated, so many times until I understood them. Thank you for your time, attention and dedication, support, always believing in me, and your anxiolytic attitude and presence as refreshing as the petrichor. You helped me overcome all kinds of challenging moments, whether related to PMMM, R code implementation, or a difficult comment from a reviewer.

Dear Frank, I remember receiving an email from you in mid-2015 inviting me to discuss the traineeship opportunity with BMS and interview scheduling, and then you called me. The manager that managed without pressing, and subtly taught me how to "EMA", how to understand clinical trial methodology and how to decode scientific advice discussions. Thank you for taking me under your wing; it was an amazing journey, and I am forever grateful and indebted to you. If you were not to select me for the traineeship, I could not imagine my life and career without the time at EMA under your mentorship and relaxed leadership and without E9(R1). Thank you for joining my PhD Supervision team and for all your time, visits to Utrecht and Nijmegen, for all the key input in my research projects, and for being there when I needed your support. We have a special relationship, and you play an important role in my development as a manager, supervisor, paranymph, and friend. Merci beaucoup!

To my PhD Supervision team as an entity. Thank you! I am so fortunate to have been your PhD student, and I hope I made you at least a little proud. I promise to carry on the estimands implementation to the best of my knowledge and effort.

Dear members of the assessment committee of my thesis and opponents: Prof. dr. A.C.G. Egberts, Prof. dr. O.H. Klungel, Prof. dr. D. Rizopoulos, Prof. dr. R.H.H. Groenwold, Dr. N. Benda, and Dr. J.B. Reitsma. I am grateful for your time and effort in assessing my thesis and for being the opponents of my PhD defence; I would like to thank you dearly for this commitment.

To all collaborators involved during my PhD and the Asterix team, it takes a village to raise a child, and multiple projects, research centres, and teams to help Marian become a methodology and statistics researcher. Thank you! Danke schön! Muchas gracias! Bedankt! I want to thank Dr. Gerd Rosenkranz, especially, for his help with the innovative rare disease methodology validation on EPARs.

I am deeply thankful to MSD and Julius Center UMCU for access to trial data for chapter 4.

Dear Rob and E9(R1) EWG, thank you for developing the estimand framework. I am grateful for your efforts and the fruits of your work; not sure if I were able of in-depth study or passionate about something other than estimands.

Dear Valentijn, I am so happy I met you and we befriended each other, I don't remember how; I just realised we are good friends. Thanks for welcoming me into your life and home and for explaining the beautiful Dutch society while we visited many museums, pubs, and restaurants. I am grateful for your time, kind words, and calming presence, always offered unconditionally. You and Saskia are amazing; Rux and I are happy and proud to call you friends. I hope you revisit us soon!

Dear Rutger, without you and your patience, I would still be coding or trying to code. It was great to share cigarettes (I am happy we both quit!). The only bus I am so glad I missed was with you. It was great to plot together our Super-Secret project that has yet to happen. I am grateful to be around you, and God bless you are you in the way you are.

To Kristoffer Segerstrøm Mørk, many thanks for your explanations and code provided in your master thesis. Hope to meet you one day to shake your hand and thank you in person. You need to know how much I benefited from your thesis and how lucky and grateful I am that I found the results of your estimand research.

To Julius Center professors and leadership teams, and Epi MSc colleagues, thank you for setting the bar high for me. I like to think I became a better methodologist because you are intelligent, hardworking, and accomplished researchers ready to share your hard-earned key learnings from your research.

Dear Coby, thank you for always being there, ready to help me, whether I needed to find a room or a warm smile & helpful attitude.

Dear Jet, thank you for being such a cool symposium co-presenter to share the mic time with and celebrate diversity with the entire Julius Center. I was happy that you were also detail-oriented, which helped us deliver smooth interventions and coordination on stage.

Dear Katrien G, LBP! Thank you for keeping me hydrated and healthy. It was fun and helpful to do the Epi MSc together, and it was an absolute pleasure to be roommates. Thank you for pushing my limits at master programme courses and HIIT boot camps after enjoying cakes across the van Geuns! You and Guido are wonderful, and I am happy I have met you.

Dear Fien, we had a great time at concerts and festivals, at yoga, around the ice cream hut, at the botanical garden, at Theehuis Rhijnauwen, at the Epi MSc courses, and pretty much everywhere. You have such a beautiful family, and I am happy I got to know your folks and friends. Thanks for induction training in the Dutch society, always being there with any information I needed, and for sharing your experience of the last months of your PhD thesis preparation. Thanks a lot for being a great colleague, roomie, and friend!

Dear Anneke, thank you for your warmth and for helping me have a head start on my PhD track. Thank you for keeping me hydrated and healthy!

Dear Alicia and Giske, thank you for your radiant presence and for being such great & happy roomies; that always helped!

Dear Julien, mon capitain, despite not working together and not speaking too often, I perceive you as a close friend. Each time we had the opportunity, we had an excellent bonding time at your PhD defence party or at EMA. I hope to keep in touch and keep the connection alive. I cannot thank you enough for sharing my CV with Kit and Rolf.

Dear Stavros, I wish we had worked more together to learn more from you. Your perspective on pretty much all possible things was always distinct, and your shared experience gave me solace and informed me more and better about the researcher's life. It was a bit worrying but quite refreshing, too, when we met at the Stratenum entrance that Saturday in 2017. I still wonder sometimes why you gifted me an ashtray when I was trying to quit smoking ☺

Dear Konstantinos, thank you for your patience, R coding lessons & on-the-job training. I am a better coder and statistician, thanks to you. You were and are an example of resourcefulness, kindness, and desire to learn more.

Dear Maarten and Rolf, we had a relatively short but excellent and productive run together. Thanks to you, I have learned to write better for loops and more efficient code, to structure my first abstract and manuscript, and that parting way does not lead to an end.

Dear Timo, thank you for getting me up to speed with our measurement error systematic review and for sharing your experience during your PhD. It was nice to spend a Christmas party at your cool World Trade Center office.

Dear Renée, thank you for always being there for me and being such a friendly & kind person in so many aspects. You have a magical personality; I do not know how you do it. Still, it was always reassuring and relaxing to discuss with you, whether it was about the PhD project, a publication, or about…pretty much anything! Thank you!

Dear Jesse, you were so many times there for everybody, a solid presence for friendships and further connections; understated, but you have my deep appreciation. Thanks a lot for being a genuine friend, for keeping me updated with F1 development and results, and for your helping me with the logistics of my PhD defence day. Inevitably, we will keep in touch!

Dragă Mircea, mă bucur că ne-am dumirit după 1 an de zile de împrumutat țigări și cerut brichete că suntem din aceeași țară. Mă bucur că păstrăm legătura, intermitent, dar o păstrăm.

Dear Arjan, you were the most fantastic alive vampire that could dance. I am so happy we partied together and thankful for all the memories we have and those we do not have because our brains did not manage to register.

Dear Rebecca, thank you for helping me to figure out the code for linear mixed effects models.

Dear Caroline, thank you for your support during the Asterix project and for involving me in teaching various courses. Your trust helped me grow so much!

Dear Cas, thank you for your patience and thoughtful teaching during my epi master courses.

Dear Bert, thank you for all your support with all possible and impossible matters and for being a role model for a healthy lifestyle!

To other colleagues from Julius Center, Paul, Peter (2x), Victor, Lotty, Rob, Thomas, Kevin, Carl, Jetske, Maria, and Hae-Won, it was an absolute pleasure working and learning from you; thank you for your support.

Rene, thanks for your help and positive attitude and for all the stats learnings from the courses you taught and for all the R coding tricks, for your support with my contract and guest account, and for being my MSc Epi project examiner, and…, and for being such a nice & kind role model!

Dear Dr. Bas Engel and Dr. Roula Tsonaka, thank you for accepting me at your mixed and longitudinal modeling course at Leiden University and welcoming me as your student. I have learned so much!

Dear Taco, thank you for all estimands scientific advice. I hope you know how valuable this learning experience was and how it fueled my passion and drove my knowledge about estimands. Bedankt!

Dear Jacqueline and Laura, thank you for sharing estimands knowledge and being kind and welcoming.

Dear Daniela, thank you for trying to get me in shape for the soon-to-be almost half-marathon and the support offered during my PhD.

Dear Antonis, ela man, you are such an amazing human, skilled in all sorts of dimensions, and such a kind person. I could not have had a better buddy to run a half-marathon quarter-marathon that pushed our limits. To the next error of this kind!

Dear Jay, it was great to be roomies. Thank you for indirectly teaching me to take things lightly. It is excellent that we keep in touch!

Dear Marjon, thank you for welcoming me at CBG-MEB and into the Dutch society and for all the support with my research, presentations, posters, and scientific communication of my PhD research and defence.

To CBG-MEB Methodology Working group. Thank you for all the learnings from scientific advice, centralised procedures, and methodological guidelines and for running the meetings in English. I miss everyone dearly!

André, thanks for the input and insights in the estimands narrative review, the positive vibe, and for inspiring the catchy title of our narrative review, “old wine in new barrels”.

To my former EMA colleagues: Veronica, Andrew, Inês, Ina, and Olivier. Thank you for accepting me on the BMS team and helping me find my way around CHMP, SAWP, working groups, and EMA ways of working in general. Your support during my traineeship was such a blessing.

To former versions of myself, thank you! You were great; I applaud you for choosing the right paths and staying away from easy, quick, or dirty, and mediocrity. I hope you are proud of yourself. I will carry on your work passionately and with a commitment to make the most of your effort and talent. Congratulations for the excellent and difficult decisions, for being bold enough to want more and better, and for working and striving to obtain that more and better, whatever the perceptions were.

To future versions of myself, I bid you good luck! Please stick to keep on learning from everyone and anyone. This you always knew best and kept your ego in check.

Ștechi, mă bucur enorm că ai fost colegul meu de cameră din timpul facultății și îți multumesc pentru toată aprecierea și pentru că ai fost mult timp cel mai mare fan al meu, a contat mai mult decât îți poți imagina. Ai o familie minunată alături de Ana și Theodora, mă bucur că vă cunoaștem și că petrecem timp împreună.

Trif și Laura, Iulian și Alexandra mă bucur foarte tare că existați, împreună, și că petrecem timp de calitate de fiecare dată când ne întâlnim, pe FaceTime sau când călătorim.

Dear Corina, Ryan, and Robert thank you for showing me different perspectives, for helping me become a better friend to myself, and for helping me become wiser at becoming happier.

To Romania, thank you for raising me and teaching me how to walk, run, count, read and write. Thank you for Ruxandra!

To the United Kingdom, thank you for teaching me that the world is larger and more beautiful than I thought.

To the Netherlands, thank you for teaching me how to cycle and how to stats.

To France, thank you for teaching me how to enjoy and find new perspectives in how I walk, cycle, read, write, or stats.

To Switzerland, thank you for teaching me to do what all the other countries taught me to do in a relaxed and mature way.

Dear BBU team, dear Biogen, thank you for welcoming me to the team, supporting me, and giving the green light to many of my initiatives. We have hit it off very nicely, and it is excellent being your colleague and a pleasure working and learning with you. I am happy you share my enthusiasm about estimands, and I look forward to many achievements in our future endeavours, executed with passion and rigour.

Rux, draga mea, mulțumesc pentru toată răbdarea, ajutorul, înțelegerea, sprijinul și pentru toată dragostea oferită. Mă bucur enorm că exiști în viața mea și mă bucur foarte tare de cum evoluăm împreună și cum învățăm unul despre și de la altul.

Dragă Mamă, dragă Tată, mulțumesc mult pentru tot efortul, sacrificiile și pentru tot sprijinul oferit de-a lungul anilor. Mă bucur nespus că sunteți părinții mei.

Dragă Anca, îți mulțumesc pentru toată răbdarea, grija și sprijinul oferite necondiționat de când eram mic. Mulțumesc pentru toate jumătățile de cornuri cu ciocolată, pentru geci, fesuri și pentru atenția și ocrotirea din adolescență, din facultate și după. Sunt foarte mândru de tine și de performanțele tale, ai fost primul meu exemplu de succes la școală, iar asta a contat enorm pentru mine. Mulțumesc mult că ești sora mea!

Dragă Bia, ești singura mea nepoțică dar nici că îmi puteam dori mai mult. Ești o bucurie în familia noastră, așa o persoană deșteaptă și blândă, sunt foarte mândru de tine și îți doresc succes în tot ce îți propui.

Dragă Dan, mulțumesc frumos pentru tot suportul oferit de-a lungul timpului, mai ales la București, în timpul facultății dar foarte mult și după.

Dragă bunică Maria, mulțumesc mult că m-ai crescut, că ai avut grijă de mine tot timpul și că m-ai ocrotit cum ai știut mai bine. Sărut-mâna!

Dear God, thank you for keeping us healthy, and I hope you grant us the power and patience to take in everything that comes our way.

About the author

Marian Mitroiu was born on the 18th of May 1988 in Buzău, România. In 2007, he finished high school at B.P. Hasdeu National College. In the same year, he started Pharmacy studies at UMF Carol Davila Faculty of Pharmacy, in Bucharest. Then, he specialized in Clinical Pharmacy, did a Pharmacovigilance MSc at Iuliu-Haţieganu University in Cluj-Napoca and grew an interest in signal detection. While working in pharmacovigilance, he discovered a new passion for biostatistics, and decided to pursue a Biostatistics degree at University of Bucharest. He did a traineeship at European Medicines Agency in London at the Biostatistics and Methodology Office, where he started learning more about regulatory statistics and drug development in the EU/EEA. He had the fortunate and unique opportunity to support the ICH E9(R1) Expert Working Group that developed the estimands addendum. In 2017, he started his doctoral studies at Julius Centre, UMC Utrecht and Utrecht University. He worked in parallel at CBG-MEB being part of the Methodology Group where he consulted for scientific advice procedures focused on estimand methodology and was statistical assessor for centralised procedure scientific assessment. He combined his estimand research with the postgraduate master programme in Epidemiology at Utrecht University, which he graduated in 2021, with a specialization in Medical Statistics. Marian joined Biogen as a biostatistician in August 2021 and continues the estimand methodology research, implementation and application in clinical drug development, being a member of various estimand-related working groups.

Authors' details

Mitroiu M[1,2], Oude Rengerink K[1,2], Teerenstra S[1,3], Pétavy F[4], Roes KCB[1,2].

[1] Methodology Working Group, College ter Beoordeling van Geneesmiddelen, Graadt van Roggenweg 500, 3531 AH Utrecht, The Netherlands.

[2] Department of Biostatistics and Research Support, Clinical Trial Methodology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands.

[3] Department for Health Evidence, Section Biostatistics, Radboud Institute for Health Sciences, Radboud University Medical Center, Geert Grooteplein 21, 6525 GA Nijmegen, The Netherlands.

[4] Biostatistics and Methodology Support Office, European Medicines Agency, Domenico Scarlattilaan 6, 1083 HS Amsterdam, The Netherlands.

Authors' note

The views expressed in this thesis are the personal views of the author(s) and may not be understood or quoted as being made on behalf of or reflecting the position of the regulatory agency/agencies or organisations with which the author(s) is/are employed/affiliated.

List of other publications:

1. Mitroiu M, Oude Rengerink K, Pontes C, Sancho A, Vives R, Pesiou S, et al. Applicability and added value of novel methods to improve drug development in rare diseases. Orphanet J Rare Dis. 2018 Dec;13(1):200.

2. Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, Smeden M van. Measurement error is often neglected in medical literature: a systematic review. Journal of Clinical Epidemiology. 2018 Jun 1;98:89–97.

3. Kersten RFMR, Öner FC, Arts MP, Mitroiu M, Roes KCB, de Gast A, et al. The SNAP Trial: 2-Year Results of a Double-Blind Multicenter Randomized Controlled Trial of a Silicon Nitride Versus a PEEK Cage in Patients After Lumbar Fusion Surgery. Global Spine Journal. 2021 Jan 7;2192568220985472.

4. Kersten RFMR, Öner FC, Arts MP, Mitroiu M, Roes KCB, de Gast A, et al. Letter to the editor regarding "Two-year results of a double-blind multicenter randomized controlled non-inferiority trial of polyetheretherketone (PEEK) versus silicon nitride spinal fusion cages in patients with symptomatic degenerative lumbar disc disorders." J Spine Surg. 2021 Jun;7(2):249–51.