CODE-EHR best-practice framework for the use of structured electronic health-care records in clinical research

Dipak Kotecha, Folkert W Asselbergs*, Stephan Achenbach, Stefan D Anker, Dan Atar, Colin Baigent, Amitava Banerjee, Birgit Beger, Gunnar Brobert, Barbara Casadei, Cinzia Ceccarelli, Martin R Cowie, Filippo Crea, Maureen Cronin, Spiros Denaxas, Andrea Derix, Donna Fitzsimons, Martin Fredriksson, Chris P Gale, Georgios V Gkoutos, Wim Goettsch, Harry Hemingway, Martin Ingvar, Adrian Jonas, Robert Kazmierski, Susanne Løgstrup, R Thomas Lumbers, Thomas F Lüscher, Paul McGreavy, Ileana L Piña, Lothar Roessig, Carl Steinbeisser, Mats Sundgren, Benoît Tyl, Ghislaine van Thiel, Kees van Bochove, Panos E Vardas, Tiago Villanueva, Marilena Vrana, Wim Weber, Franz Weidinger, Stephan Windecker, Angela Wood, Diederick E Grobbee, on behalf of the Innovative Medicines Initiative BigData@Heart Consortium, European Society of Cardiology, and CODE-EHR International Consensus Group*

Big data is important to new developments in global clinical science that aim to improve the lives of patients. Technological advances have led to the regular use of structured electronic health-care records with the potential to address key deficits in clinical evidence that could improve patient care. The COVID-19 pandemic has shown this potential in big data and related analytics but has also revealed important limitations. Data verification, data validation, data privacy, and a mandate from the public to conduct research are important challenges to effective use of routine health-care data. The European Society of Cardiology and the BigData@Heart consortium have brought together a range of international stakeholders, including representation from patients, clinicians, scientists, regulators, journal editors, and industry members. In this Review, we propose the CODE-EHR minimum standards framework to be used by researchers and clinicians to improve the design of studies and enhance transparency of study methods. The CODE-EHR framework aims to develop robust and effective utilisation of health-care data for research purposes.

Lancet Digit Health 2022; 4: e757–64

Published Online August 29, 2022 https://doi.org/10.1016/S2589-7500(22)00151-0

*Joint first authors

Institute of Cardiovascular Sciences (Prof D Kotecha PhD) **and Institute of Cancer and Genomic Sciences** (G V Gkoutos PhD)**, University of Birmingham, Birmingham, UK; Health Data Research UK Midlands, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK** (Prof D Kotecha, G V Gkoutos)**; Health Data Research UK London, London, UK** (Prof F W Asselbergs MD PhD, Prof A Banerjee DPhil, Prof S Denaxas PhD, Prof H Hemingway MBBS, R T Lumbers PhD)**; Department of Cardiology, Division of Heart and Lungs** (Prof D Kotecha)**, and Department of Epidemiology** (Prof D E Grobbee PhD)**, Julius Center for Health Sciences and Primary Care** (G van Thiel PhD)**, University Medical Centre Utrecht, and Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences** (W Goettsch PhD)**, University of Utrecht, Utrecht, Netherlands; Institute of Cardiovascular Science and Institute of Health Informatics, Faculty of Population Health Sciences, University College London, London, UK** (Prof F W Asselbergs)**; Department of Cardiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany** (Prof S Achenbach MD)**; Department of Cardiology and Berlin Institute of Health Centre for Regenerative Therapies, German Centre for Cardiovascular Research,**

Introduction

In the context of ageing populations and increasing multimorbidity in all disease areas,[1–3] large-scale, real-world data provide an opportunity to improve understanding of the epidemiology of rare and common conditions and to improve prevention strategies and treatment stratification for these conditions.[4] Specific management for individual patients is essential to reduce health-care costs and provide patient-centred care that can improve both the quality of life and prognosis of a patient. Use of controlled trials in real-world settings, either within registries or routine clinical practice, is now possible and could provide more generalisable results to the population at large than conventional studies which are often based on selected cohorts.[5]

Health-data science has undergone rapid development in the past decade, including the common use of electronic health-care record (EHR) systems that condense clinical episodes into coded, structured labels for diseases and health-care utilisation.[6] However, concerns about the quality, data privacy, transparency, and comparability of these systems have restricted the use of evidence generated with structured health-care data. These concerns have also restricted acceptance of evidence generated with structured health-care data by regulators, reimbursement authorities, and guideline task forces. Despite the availability of numerous reporting standards, no consensus regarding how to realise the Findable, Accessible, Interoperable, and Reusable (FAIR) principles[7] in the context of structured health-care data has been established. Existing reporting checklists ask authors to indicate the section of their paper in which particular design issues have been discussed; for example, Strengthening the Reporting of Observational Studies in Epidemiology (known as STROBE) for observational studies,[8] Reporting of Studies Conducted Using Observational Routinely-Collected Health Data (known as RECORD) for routinely collected health data,[9] and Consolidated Standards of Reporting Trials-Artificial Intelligence (known as CONSORT) for artificial intelligence interventions.[10] However, these checklists are often long, no minimum standards are specified, and adherence does not relate to study quality or even the quality of transparency for that domain.[11] Although checklists can benefit research quality, they are often used as a formality to facilitate journal publication. In a study of radiology journals, only 15% (120/821) of authors used the reporting guideline when designing their study.[12] With a proliferation of reporting checklists for every scenario, authors and readers are increasingly confused about the value of these checklists. As of Feb 14, 2022, 488 reporting checklists were registered with Enhancing the Quality and Transparency of Health Research (EQUATOR) and 111 reporting checklists were in development.

In observational and randomised clinical research that uses EHRs and other structured data, the source of data, its manipulation, and data governance are of crucial importance to extrapolating results. Clarity from the stakeholders is needed to provide a quality framework to enhance the design and application of clinical research that increasingly depends on these new sources of data. This Review shows the collaboration of a wide range of international stakeholders with the task of improving the use of routinely collected health-care data. The CODE-EHR framework was coordinated by the European Society of Cardiology (ESC), a non-profit organisation of health-care professionals, and the BigData@Heart consortium, a public–private partnership funded by the

Charité Universitätsmedizin, Berlin, Germany (Prof S D Anker MD PhD); **Department of Cardiology, Oslo University Hospital, Oslo, Norway** (Prof D Atar MD PhD); **Institute of Clinical Medicine, University of Oslo, Oslo, Norway** (Prof D Atar); **Medical Research Council Population Health Research Unit, Nuffield Department of Population Health, Oxford, UK** (Prof C Baigent MSc); **Clinical Trial Service Unit and Epidemiological Studies Unit** (Prof C Baigent) **and European Heart Journal, Oxford University Press** (F Crea MD), **University of Oxford, Oxford, UK; University College London Hospitals NHS Trust, London, UK** (Prof A Banerjee); **European Heart Network, Brussels, Belgium** (B Beger MA, S Løgstrup MA, M Vrana MA); **Bayer, Stockholm, Sweden** (G Brobert PhD); **Division of Cardiovascular Medicine, John Radcliffe Hospital, University of Oxford National**

European Union Innovative Medicines Initiative. We aim to explain opportunities and limitations of using structured health-care data in research and develop a framework for a broad audience of global stakeholders across all disease areas. The CODE-EHR framework seeks to leverage the digitisation of health data to increase the efficiency of health-care systems and improve the lives and wellbeing of patients.

Stakeholder development of the CODE-EHR framework

A full range of stakeholders participated in developing the CODE-EHR framework, including regulators (US Food and Drug Administration and European Medicines Agency), governmental agencies (European Commission, the UK National Institute for Health and Care Excellence, and Innovative Medicines Initiative), medical journals (*The BMJ*, *European Heart Journal*, *The Lancet*, and *The Lancet Digital Health*), patient advocacy groups (European Heart Network and ESC Patient Forum), representatives from the pharmaceutical industry, health-care payers, academic institutions, and professional societies. The CODE-EHR framework was initially developed in two stakeholder meetings (July 7, 2020, and Oct 26, 2020) consisting of presentations from experts in a specific medical or technological field, with smaller group sessions and plenary sessions to formulate statements on important topic areas. An iterative process with virtual work (work via email and contribution to shared online workspaces) was used to achieve a consensus, with a further meeting on March 10, 2022, to finalise the checklist.

We aimed to develop pragmatic advice for the use of structured health-care data in trials and observational studies that is not dependent on particular diseases and that meets the expectations of stakeholders and the general public. Our objectives were to provide guidance for this increasingly important field in medicine, thereby enhancing the value of regularly collected data to improve future patient wellbeing. Detailed text on the current stage of technological development for research using health-care data, in addition to important challenges and limitations of using these data, was developed by the stakeholders and supported by a writing committee. We aimed to explain the need for common standards when using health-care data (eg, coding systems and linkage), their use in all medical areas, and to show how a social licence from the public can lead to the cocreation of research with a public health benefit (appendix 1). To improve clinical practice, there are important challenges that should be overcome in all areas of structured health-care data (figure 1).

The output of the stakeholder meetings and iterative discussions was condensed into four themes: technical process and data stewardship; data security and privacy; publications using structured health-care data; and the needs of regulators, reimbursement authorities, and clinical practice guidelines. Important statements and advisories from the meetings have been summarised (panel).

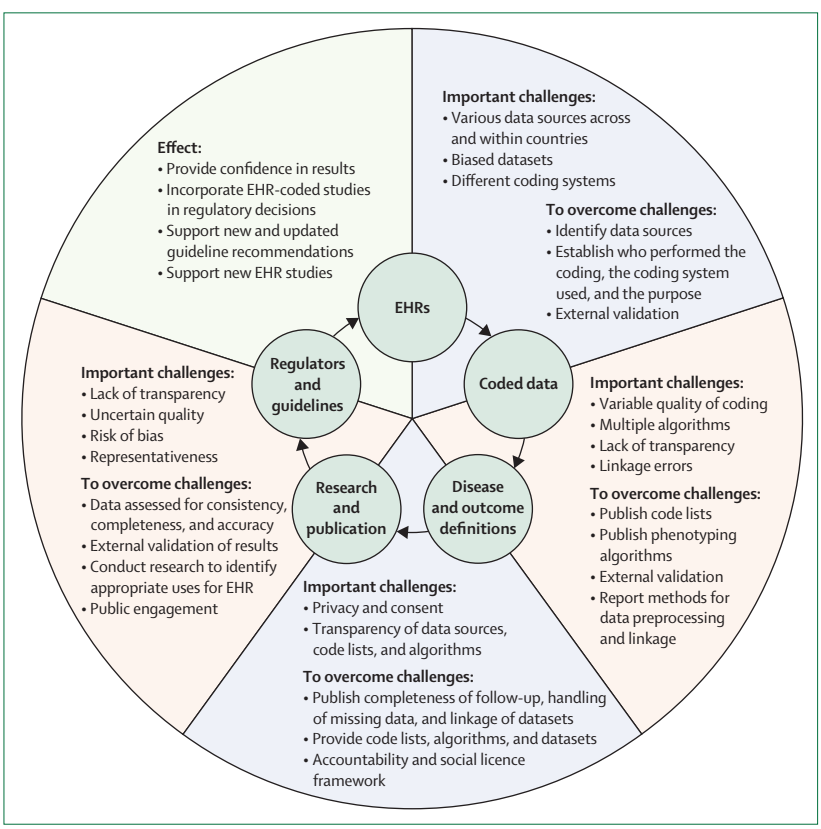

Figure 1: **From structured health-care data to improved patient care**
Key challenges and the paths to improvement leading to sustainable impact from EHR-based research studies. EHR=electronic health-care record.

Patient and public involvement

The CODE-EHR consensus approach has benefited from patient and public involvement throughout the development process, including representation from the ESC Patient Council and the European Heart Network, an alliance of foundations and associations supporting patients and representing patient interests. Methods for engaging the public in future research can constructively benefit research using big data (figure 2).

CODE-EHR reporting framework

The process from structured health-care data to clinical research output is complex. To support further development of effective clinical research in a transparent way, stakeholder delegates established a consensus of the need for a set of minimum standards that authors could use to enhance the design of a study, reporting of results, and research output. The CODE-EHR Minimum Standards Framework allows authors to report how structured health-care data were used in their research study (either in patient identification, disease phenotyping, or outcome

Panel: Output from the stakeholder consensus meetings

Key consensus statements and advisories

(1) Technical process and data stewardship

Research using structured health-care data conducted according to the Findable, Accessible, Interoperable, and Reusable [or FAIR] data principles of findability, accessibility, interoperability, and reusability:[13]

- Important considerations are transparency of who performed the coding, the coding system used, and the purpose of coding (eg, reimbursement and diagnosis).

Clear and consistent identification and description of the sources of electronic health-care record (EHR) data:

- Code lists and phenotyping algorithms can be described in detail and published, ideally before a study commences (eg, on a coding repository or open-source archive). The minimum data required to meet the definitions will depend on the use case and can be reported to enhance transparency, in addition to the rationale for why certain decisions were made (eg, why one code was chosen over another, or what the effect would be if data collection periods were changed).

Validation at local, regional, and global levels:

- Evidence showing how algorithms have been externally validated, and also what quality assessment was performed on the research findings; for example, on the accuracy, completeness, and timeliness of the data.[14] Data quality rules can be used to assess coded data and allow comparisons across institutions and countries.[15]

Reporting of methods used for data preprocessing and data linkage:

- This includes the methods used to assess the quality of linkage and the results of any data preprocessing and linkage (with provision of false positive and false negative rates, comparisons of linked and unlinked data, and any sensitivity analyses).[16] A flow diagram showing the processes for cleaning and linking different coding sources and datasets can aid understanding of the study design.

Reporting of the governance framework underpinning the study from a technical or data stewardship standpoint:

- This includes a clear purpose for data gathering and the parameters and time limit of consent, clear mechanisms for data processing ("what happens with my data"), and a description of what the data can and cannot be used for (ie, the mandate given for research).

(2) Data security and privacy

Working towards a new, sustainable mandate from the public and patients to use their health data may require moving away from abstract rules and regulations and towards more constructive governance, in which trust is a central concept:

- The trust of patients and the public in research institutions and in science is pivotal because of the liberties they give to researchers to use their data, which are the product of a social licence based on this trust.

Gaining this trust would benefit from understanding what society and stakeholders expect from scientists conducting health data research, with engagement of stakeholders from the concept stage:

- Cocreation of data governance based on inclusion of patient and public communities and dialogue with researchers is crucial for ethical and sustainable governance, and to translate expectations into scientific research and scientific output.

Researchers and big data consortia have to be mindful that trustworthiness comes with the duty to act in ethically responsible ways:

- This concerns two areas; first is the competence in data handling (meaning that systems are in place to ensure data protection and there is a framework of rules and regulations for data sharing), and second is what motivates the data analysis. Ongoing dialogue can ensure that public values continue to be aligned with the governance structures of health data research projects. Questions arise as to how to measure success at implementing public values into research, and what levels of public support are sufficient to grant a mandate for data usage.

Complex organisational structures may be less important for this trust than is often asserted:

- Complicated rules and regulations may do more harm than good in establishing the conditions for public trust in big data health research to flourish, and as a result be counterproductive especially when a social licence has not been adequately achieved.[17]

Embracing values such as transparency, reciprocity, inclusivity, and service to the common good:

- These values can be embedded into the governance framework of big data health research.[18] This calls for constructing a narrative that researchers and research consortia can be held accountable so that patients and the wider public are willing, and consistently willing, to place their trust in health research projects.

Governance could be aided by developing a framework for accountability:

- This includes clear distinctions between anonymised, pseudonymised, and aggregate data along with plain language explanation to participants and users, and discrimination between primary and secondary use of data sources.

(3) Publications using structured health-care data

Accountability for the source of data and how the data have been collected (traceability):

- As with data security, a framework of accountability would enable editorial teams in medical journals to be aware of the technical processes before data analysis.

Sharing of data, codes, and algorithms used to analyse datasets:

- Similar to the requirement for preregistration of clinical trials and prepublication of protocols, journals could restrict publication where the coding within a study is not shared.

(Continues on next page)

Institute for Health and Care Research Oxford Biomedical Research Centre, Oxford, UK (Prof B Casadei MD DPhil); **European Society of Cardiology, Sophia Antipolis, France** (C Ceccarelli BA); **Royal Brompton Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK** (Prof M R Cowie MD); **School of Cardiovascular Medicine Sciences, King's College London, London, UK** (Prof M R Cowie); **Department of Cardiovascular and Pulmonary Sciences, Catholic University of the Sacred Heart, Rome, Italy** (F Crea); **Vifor Pharma, Glattbrugg, Switzerland** (M Cronin MD PhD); **Ava, Zurich, Switzerland** (M Cronin); **Alan Turing Institute, London, UK** (Prof S Denaxas); **British Heart Foundation Data Science Centre, London, UK** (Prof S Denaxas); **Bayer, Leverkusen, Germany** (A Derix PhD, L Roessig MD, C Steinbeisser MA); **School of Nursing and Midwifery, Queen's University Belfast, Northern Ireland** (Prof D Fitzsimons PhD); **Late Clinical Development, Cardiovascular, Renal and Metabolism** (M Fredriksson MD), **and Data Science and Artificial Intelligence** (M Sundgren PhD), **Biopharmaceuticals, AstraZeneca, Gothenburg, Sweden; Leeds Institute of Cardiovascular and Metabolic Medicine and Leeds Institute for Data Analytics, University of Leeds, Leeds, UK** (Prof C P Gale PhD); **Department of Cardiology, Leeds Teaching Hospitals NHS Trust, Leeds, UK** (Prof C P Gale); **National Health Care Institute, Diemen, Netherlands** (W Goettsch); **Department of Clinical Neuroscience, Karolinska Institutet, Solna, Sweden** (Prof M Ingvar MD PhD); **Department of Neuroradiology, Karolinska University Hospital Stockholm, Stockholm, Sweden** (Prof M Ingvar); **Data and Analytics Group, National Institute for Health and Care Excellence, London, UK** (A Jonas MSc); **Office of Cardiovascular Devices** (R Kazmierski PhD) **and Centre for Devices and Radiological Health** (Prof I L Piña MD MPH),

US Food and Drug Administration, Silver Spring, MD, USA; Institute of Health Informatics, Barts Health NHS Trust and University College London Hospitals NHS Trust, London, UK (R T Lumbers); **Centre for Molecular Cardiology, University of Zurich, Zurich, Switzerland** (Prof T F Lüscher MD); **Research, Education and Development, Royal Brompton and Harefield Hospitals, London, UK** (Prof T F Lüscher); **Faculty of Medicine, Imperial College London, London, UK** (Prof T F Lüscher); **European Society of Cardiology Patient Forum, European Society of Cardiology, Brussels, Belgium** (P McGreavy MSc); **College of Medicine, Central Michigan University, Midlands MI, USA** (Prof I L Piña); **Steinbeisser Project Management, Munich, Germany** (C Steinbeisser); **Centre for Therapeutic Innovation, Cardiovascular and Metabolic Disease, Institut de Recherches Internationales Servier, Suresnes, France** (B Tyl MD); **The Hyve, Utrecht, Netherlands** (K van Bochove MSc); **Hygeia, Mitera, Hospitals Hellenic Health Group, Athens, Greece** (Prof P E Vardas PhD); **European Heart Agency, European Society of Cardiology, Brussels, Belgium** (Prof P E Vardas); **The British Medical Journal, London, UK** (T Villanueva MD, W Weber MD PhD); **Rudolfstiftung Hospital, Vienna, Austria** (Prof F Weidinger MD); **Department of Cardiology, Inselspital, University Hospital Bern, Bern, Switzerland** (Prof S Windecker MD); **Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK** (Prof A Wood PhD)

Correspondence to: Prof Dipak Kotecha, Institute of Cardiovascular Sciences, University of Birmingham, Birmingham B15 2TT, UK
d.kotecha@bham.ac.uk

See **Online** for appendices 1–5

(Panel continued from previous page)

Demonstration of data validity and robust analysis:
- The US Food and Drug Administration and European Medicines Agency already suggest independent checking or accreditation of data sources; this accreditation could be provided to editors to increase their confidence in data quality.

Balancing the speed of publication against requirements for data validation:
- Prompt publication (eg, of results with immediate public health implication) needs to be balanced against validation of data sources to ensure authenticity.

Scientific advice committees with experts in big data analytics to aid journal editorial teams:
- The skill set required in editors and reviewers for studies using structured health-care data is not the same as having statistical or clinical trials experience; expertise in EHR data and respective coding systems could add value to the journal review process.

Widening gap between the knowledge of physicians and the advanced methodologies used in big data papers:
- Medical students, graduate students, and practising clinicians, as well as hospital managers and leadership, need training in health data management and analysis. This is important to build a digital workforce with increased capacity and capability to translate publications using new approaches to improve patient care.[19]

(4) Addressing the needs of regulators, reimbursement authorities, and clinical practice guidelines

EHR-based trials have the potential to generate reliable and cost-efficient results:
- Each type of trial and each type of clinical question is considered in an individual context, including under what circumstances a particular type of EHR process could assist in answering questions about a particular intervention, and with what limitations.

Further research may help explore cases in which EHR studies produce valuable evidence, and when they might be flawed:
- This will generate confidence in regulators for future EHR studies, and for guideline task forces to appropriately appraise evidence.

Quality standards will help to ensure that the information recorded in EHR systems represents real events without bias:
- This will enable confidence that trials using EHRs can produce reliable results on efficacy and safety, and could include examination of the validity of both data sources and data analyses.

Source data validation to report on appropriate computational phenotypes:
- This could be supported by an independent adjudication committee to examine a subset of the EHR and confirm outcome events. The use of artificial intelligence techniques could facilitate larger validation studies by automated extraction of supporting text from clinical notations. Such validation exercises can be preregistered; for example, in the form of a Study-Within-A-Trial.[20] Another possibility is for researchers to provide consented and anonymised gold standard cases to benchmark against, or for data from devices used to verify codes (such as lead fractures). The value of synthetic datasets for validation, which mimic real data, needs further exploration.

Mixed-model approaches to collect data on particular endpoints:
- May be valuable for situations where the EHR does not reliably collect relevant data; for example, where patients or clinicians, or both, are asked for information, or data are collected via wearable devices or telemonitoring. In some cases, parallel monitoring of patients alongside the EHR study may provide additional confidence (eg, to identify serious unexpected adverse events). Technological advances in EHR systems will help, such as the ability to retrieve EHR data on a daily basis to support clinical trials.[21]

Taking advantage of the many real-world data initiatives to support new research:
- Government agencies, regulators, charities, and professional bodies have initiated programmes for better use of real-world data that can support further activity and dissemination.

derivation). Preferred standards indicate a high attainment of quality and can be used to improve the future trajectory of research. The checklist (figure 3) was created through an iterative process on the basis of the stakeholder proposals and includes five important areas of enhanced transparency: how and why coding was performed; the process of constructing and linking datasets; clear definitions of both diseases and outcomes; the approach to analysis, including any computational methods; and showing good data governance.

The CODE-EHR framework aims to improve the quality of studies using structured health-care data and provide confidence in their use for clinical decision making. A step-by-step approach to completion of the CODE-EHR reporting checklist, with relevant best-practice examples, is provided (appendix 2). A detailed description of the workflow that led to the checklist is also available (appendix 3). The checklist is available as Word document (appendix 4) and PDF (appendix 5) versions.

Discussion

Technological progress has led to rapid progress in heath data systems, with immediate effects in daily clinical practice. The potential for improving patient care and outcomes is clear, as are the challenges and limitations.[24] Big data analytics now support large-scale and

cost-effective clinical research, with trials based within registries or the EHR now indicating a new stage of evidence generation. The processes of using routinely collected heath-care data can be further developed by an accompanying social licence from the public and upskilling knowledge for all stakeholders. Cocreation of research studies and shared decision making with patients and the public[25] is an important way to ensure appropriate data stewardship and privacy, leading to clinical effects through robust publications, regulatory decision making, and practice guidelines. In this Review, we reported a global multistakeholder process to develop a framework for researchers to use in the design and reporting of studies that include structured or coded health-care data.

Digital health records are confusing for most researchers, with varying access to a multitude of different coding systems and classifications and considerable differences between coding systems across and within countries. Linkage of different health sources is often a crucial component of research based on structured health-care data, but this aspect is frequently overlooked when reporting such studies. Data privacy and the licence for research can be severely compromised if linkage is not secure. Therefore, our focus is on transparency about how data are coded and linked and how these approaches undertaken by researchers are openly discussed and documented. The stakeholder consensus meetings highlighted clarity of methods as an important concern for future research, supported by evidence that few studies provide sufficient detail to understand the research process.[26,27] The emergence of registry-based and EHR-based randomised controlled trials (NCT04700826; NCT01093404)[21,28] reinforces the imperative for improvements to define new concepts for quality research. With the development of robust analytics supported by machine learning algorithms,[29] similar approaches have already been used to support artificial intelligence in health care.[30]

A lack of transparency has a direct effect on the value of research using coded records, with issues for medical journals, regulators, clinical guideline writers, clinicians, and the public. Bringing together the full range of these stakeholders, we aimed to take full advantage of technical developments in the past 10 years to use structured health-care data in research, to handle limitations directly, and to provide a framework across all medical fields in which coded data can be used to improve patient care. Several other overlapping themes emerged from the discussions, including the generation and retainment of public trust and confidence and the need for coherent plans to deal with data security failures. Forethought about dealing with the harmonisation of data and the requirement for embedded validation methods were highlighted as important factors for future successful research. Similarly, education and communication are crucial for patients, the public, and health-care professionals to effectively use the results from structured health-care data studies.

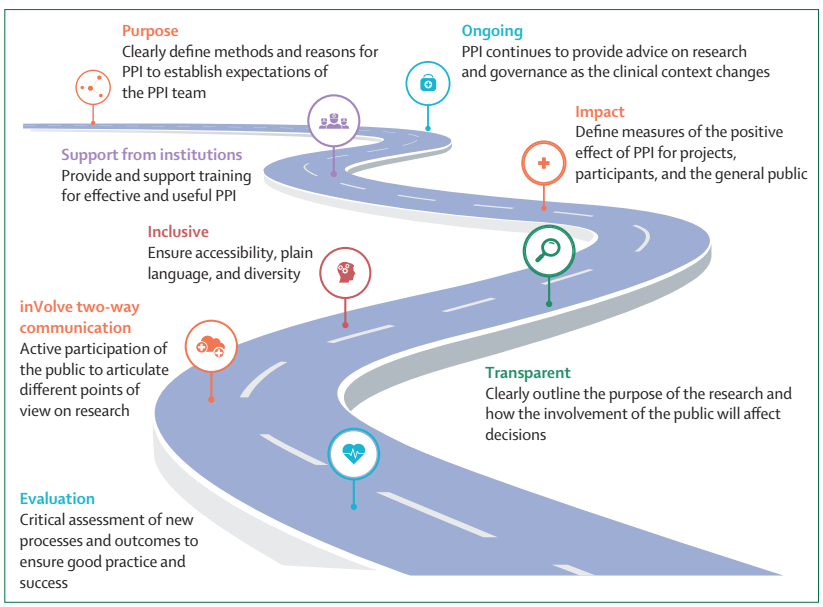

Figure 2: **PPI to improve clinical research using the POSITIVE approach**
POSITIVE steps leading to cocreation with patients and the public and improved research using big data sources. Content adapted from the Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research,[22] as used in the DaRe2THINK trial programme (NCT04700826). Adapted from Bunting and colleagues,[23] by permission of the European Society of Cardiology. PPI=patient and public involvement.

The COVID-19 pandemic has shown the need for rapid access to routine health-care data to guide and monitor clinical care and the need for a clinical trial infrastructure to allow for immediate deployment of trials in clinical practice. The digitalisation of health care, in particular the use of EHRs, offered the clinical community a unique opportunity to develop a learning health-care system that could efficiently address the effects of COVID-19. For example, information about the relationship between COVID-19 and cardiovascular disease was provided by linked EHR data that combined primary care data, hospital data, death records, and COVID-19 testing in more than 54 million people.[31] However, the COVID-19 pandemic also showed the limitations of various systems that restricted the sharing of data in real time that could direct care and help design clinical trials. Established governance of health-care data studies, security, interoperability (system architecture that spans different EHR systems and health-care providers), and phenotype definition restricted access to routine EHR data, especially during the early phases of the pandemic in each country.

The CODE-EHR framework is intended to complement available reporting checklists.[32–35] Although existing checklists are aimed at transparency in the reporting of important methodological components of clinical research, the CODE-EHR framework is designed to ensure that a common set of minimum standards are applied across all research using structured health-care data, including observational studies and controlled trials. The standards are hoped to improve research design for

all future EHR studies. Additionally, the framework supports the wider implementation of good-quality, real-world data research based on the FAIR principles.[13] Researchers are advised to use the CODE-EHR checklist in the design phase of their study to ensure that important criteria for successful research and research impact are

Date of completion:		Study name:		
Item	Objective	Framework standards	Minimum information to provide	Lead author acknowledgment
1. Dataset construction and linkage	To provide an understanding of how the structured health-care data were identified and used	Minimum: Flow diagram of datasets used in the study, and description of the processes and directionality of any linkage performed, published within the research report or supplementary documents Preferred: Provided within a prepublished protocol or open access document	(a) State the source of any datasets used (b) Comment on how the observed and any missing data were identified and addressed, and the proportion observed for each variable (c) Provide data on completeness of follow-up (d) For linked datasets, specify how linkage was performed and the quality of linkage methods	Choose one from: (1) Minimum standard not met (2) Minimum standard met or (3) Preferred standard met
2. Data fit for purpose	To ensure transparency with the approach taken, with respect to coding of the structured health-care data	Minimum: Clear unambiguous statements on the process of coding in the methods section of the research report Preferred: Provided within a prepublished protocol or open access document	(a) Confirm origin, clinical processes, and the purpose of data (b) Specify coding systems, clinical terminologies, or classification used and their versions, and any manipulation of the coded data (c) Provide detail on quality assessment for data capture (d) Outline potential sources of bias	Choose one from: (1) Minimum standard not met (2) Minimum standard met or (3) Preferred standard met
3. Disease and outcome definitions	To fully detail how conditions and outcome events were defined, allowing other researchers to identify errors and repeat the process in other datasets	Minimum: State what codes were used to define diseases, treatments, conditions, and outcomes prior to statistical analysis, including those relating to patient identification, therapy, procedures, comorbidities, and components of any composite endpoints Preferred: Provided within a prepublished protocol or open access document prior to statistical analysis	(a) Detailed lists of codes used for each aspect of the study (b) Date of publication and access details for the coding manual (please add to box below) (c) Provide definitions, implementation logic and validation of any phenotyping algorithms used (d) Specify any processes used to validate the coding scheme or reference to prior work	Choose one from: (1) Minimum standard not met (2) Minimum standard met or (3) Preferred standard met
4. Analysis	To fully detail how outcome events were analysed and allow independent assessment of the authenticity of study findings	Minimum: Describe the process used to analyse study outcomes, including statistical methods and use of any machine learning or algorithmic approaches Preferred: Provide a statistical analysis plan as a supplementary file, locked before analyses commencing	(a) Provide details on all statistical methods used (b) Provide links to any machine code or algorithms used in the analysis, preferably as open source (c) Specify the processes of testing assumptions, assessing model fit and any internal validation (d) Specify how generalisability of results was assessed, the replication of findings in other datasets, or any external validation	Choose one from: (1) Minimum standard not met (2) Minimum standard met or (3) Preferred standard met
5. Ethics and governance	To provide patients, who might or might not have given consent, and regulatory authorities the ability to interrogate the security and provenance of the data	Minimum: Clear unambiguous statements on how the principles of Good Clinical Practice and Data Protection will be or were met, provided in the methods section of the research report Preferred: Provided within a prepublished protocol or open access document with evidence of patient and public engagement	(a) State how informed consent was acquired, or governance if no patient consent (b) Specify how data privacy was protected in the collection and storage of data (c) Detail what steps were taken for patient and public involvement in the research study (d) Provide information on where anonymised source data or code can be obtained for verification and further research	Choose one from: (1) Minimum standard not met (2) Minimum standard met or (3) Preferred standard met
6. Coding manual	DOI of publication or website address: Date published:			
7. Comments				
8. Summary declaration	Choose one from: one or more minimum standards not met or all minimum standards met Number of preferred standards met: /5			

Figure 3: **CODE-EHR framework: best-practice checklist to report on the use of structured electronic health-care records in clinical research**

Directions for use for the research team: to complete the checklist, authors will need to consider these points during the design of the research to ensure that coding protocols and coding manuals are prepublished. Where applicable, it is advisable that all five minimum standards are met for an individual research study, whether observational or a controlled trial. If any component is not applicable to the study, the corresponding author can indicate why this is the case in the comment box. This checklist can accompany the article as a supplementary file on submission to the journal, with the ability for readers to review responses. A comment on the meeting of standards in the text of the method section is suggested (eg, "this study meets all five of the CODE-EHR minimum framework standards for the use of structured health-care data in clinical research, with two out of five standards meeting preferred criteria [add reference to this CODE-EHR paper; https://doi.org/10.1136/bmj-2021–069048]" or "this study meets four out of five of the CODE-EHR minimum framework standards for the use of structured health-care data in clinical research; one of the five minimum standards was not met as coding schemes were not specified prior to analysis [add reference to this CODE-EHR paper; https://doi.org/10.1136/bmj-2021–069048]"). Easy to complete form versions of this checklist are available as a Word document (appendix 4) or a PDF (appendix 5), and at https://www.escardio.org/bigdata. Directions for use for the research appraisers (patients, clinicians, regulators, and guideline task forces): where applicable, it is advisable that all five minimum standards are met for the research study to be considered robust.

being used. This process will help journal editors, regulators, guideline writers, clinicians, and patients to better appreciate the value and limitations of a study. Dissemination plans for CODE-EHR include discussions with journals to request that authors complete the checklist when submitting relevant research, attaining full registration with the EQUATOR network,[36] and using international digital health groups to engage their members and other relevant stakeholder organisations to use the checklist. After publication, the CODE-EHR framework will undergo a 2-year evaluation, including discussions with researchers using the approach, with a plan for iterative improvements to adapt to the rapidly developing field of medical research.

Conclusion

The CODE-EHR framework was designed by a multistakeholder group to improve the design and reporting of research studies using structured electronic health-care data. Research using these data is a crucial component of future health-care evaluation and administration and will have an increasingly important role in decisions regarding patient care made by regulatory, governmental, and health-care agencies in every medical specialty. The CODE-EHR checklist asks for clarity on reporting and defines a set of minimum and preferred standards for the processes involved in coding, dataset construction and linkage, disease and outcome definitions, analysis, and research governance. The CODE-EHR framework is expected to enhance research quality and value and to improve research impact using regularly collected health-care data.

Contributors

DK and FWA are joint first authors with equal contribution. Each subsequent listed author was involved in the CODE-EHR consensus process and wrote or revised sections of text and revised the final manuscript for intellectual content. All listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Declaration of interests

The BigData@Heart project has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking (grant 116074) which receives support from the Horizon 2020 research and innovation programme of the European Union and the European Federation of Pharmaceutical Industries and Associations. The funders had no role in considering the study design; in the collection, analysis, or interpretation of data; in the writing of the report; or in the decision to submit the Review for publication. DK reports grants from the European Union and the European Federation of Pharmaceutical Industries and Associations Innovative Medicines Initiative (116074), the National Institute for Health Research (NIHR CDF-2015-08-074 RATE-AF; NIHR130280 DaRe2THINK; and NIHR132974 D2T-NeuroVascular), the British Heart Foundation (PG/17/55/33087, AA/18/2/34218 and FS/CDRF/21/21032), and the ESC. DK is supported by educational grants from Boehringer Ingelheim, Bristol Myers Squibb (BMS)-Pfizer Alliance, Bayer, Daiichi Sankyo, Boston Scientific, the NIHR and University of Oxford Biomedical Research Centre, the British Heart Foundation and University of Birmingham Accelerator Award (STEEER-AF NCT04396418), Amomed Pharma, Istituto Di Ricovero e Cura a Carattere Scientifico San Raffaele, and Menarini (NCT0083244). DK receives advisory board personal fees from Bayer, Amomed, Protherics Medicines Development, and Myokardia. FWA reports grants from IMI BigData@Heart. SDA reports grants and personal fees from Vifor Pharma, Abbott, and Abbott Vascular; and personal fees from Bayer, Boehringer Ingelheim, Servier, Cardiac Dimensions, Actimed, AstraZeneca, Amgen, Bioventrix, Janssen, Respicardia, V-Wave, Brahms, Cordio, and Occlutech. CB reports grants from Medical Research Council, Boehringer Ingelheim, and NIHR. AB reports grants from AstraZeneca. BB reports grants from the European Commission. GB reports grants from the European Commission (IMI project support) and Bayer. BC reports non-financial support from Roche Diagnostics and iRhythm. CC reports grants from the European Commission. MRC reports personal fees from AstraZeneca. FC reports personal fees from Amgen, AstraZeneca, Servier, and BMS; and advisory board fees from GlyCardial Diagnostics. MC reports personal fees from Vifor Pharma. AD reports salary from Bayer. MF reports salary from AstraZeneca. CPG reports personal fees from AstraZeneca, Amgen, Bayer, Boehringer Ingelheim, Daiichi Sankyo, Vifor Pharma, Menarini, Wondr Medical, Raisio Group, and Oxford University Press; and grants from BMS, Abbott, British Heart Foundation, NIHR, and ESC. MI reports grants from World Economic Forum, Swedish Innovation Agency, and the European Commission; and collaboration with Frisq. SL reports grants from the European Commission. TFL reports grants from Abbott, Amgen, Novartis, Boehringer Ingelheim, Servier, Vifor Pharma, Sanofi, and AstraZeneca; and personal fees from Daichi Sankyo, Pfizer, and Menarini. LR reports salary from Bayer. CS reports personal fees from Bayer. BT reports personal fees from Servier. GvT reports grants from IMI. KvB reports grants from IMI BigData@Heart. PEV reports personal fees from Hygeia Hospitals Group, Hellenic Healthcare Group, ESC, and Servier. TV is working as an editor at The BMJ and Acta Médica Portuguesa and is Vice President of the European Union of General Practitioners. MV reports grants from the European Commission. SW reports grants from Abbott, Amgen, AstraZeneca, BMS, Bayer, Biotronik, Boston Scientific, Cardinal Health, CardioValve, CSL Behring, Daiichi Sankyo, Edwards Lifesciences, Guerbet, InfraRedx, Johnson & Johnson, Medicure, Medtronic, Novartis, Polares, OrPha Suisse, Pfizer, Regeneron, Sanofi-Aventis, Sinomed, Terumo, and V-Wave. SW serves as an unpaid advisory board member or unpaid member of the steering or executive group of trials funded by Abbott, Abiomed, Amgen, AstraZeneca, BMS, Boston Scientific, Biotronik, Cardiovalve, Edwards Lifesciences, MedAlliance, Medtronic, Novartis, Polares, Sinomed, V-Wave, and Xeltis. He has not received personal payments by pharmaceutical companies or device manufacturers. SW is a member of the steering or executive committee group of several investigator-initiated trials that receive funding by industry without effect on his personal remuneration. SW is an unpaid member of the Pfizer Research Award selection committee in Switzerland and of the Women as One Awards Committee. SW is a member of the Clinical Study Group of the Deutsches Zentrum für Herz Kreislauf-Forschung and of the Advisory Board of the Australian Victorian Heart Institute. He is Chairperson of the ESC Congress Program Committee, former Chairperson of the ESC Clinical Practice Guidelines Committee, and Deputy Editor of the *Journal of the American College of Cardiology Cardiovascular Interventions*. All other authors declare no competing interests.

Acknowledgments

We are grateful for the work of the CODE-EHR international consensus. The following were delegates and attendees of the meetings. Regulatory agencies: European Medicines Agency, Netherlands (Xavier Kurz); US Food and Drug Administration, USA (John Concato, Robert Kazmierski, Jose Pablo Morales, and Ileana Piña). Reimbursement authorities: National Health Care Institute, Netherlands (Wim Goettsch); National Institute for Health and Care Excellence, UK (Adrian Jonas); Dental and Pharmaceutical Benefits Agency (TLV), Sweden (Niklas Hedberg). Medical Journal Editors: *European Heart Journal*, Switzerland (Filippo Crea and Thomas F Lüscher); *The BMJ*, UK (Wim Weber and Tiago Villanueva). Industry representatives: AstraZeneca (Martin Fredriksson and Mats Sundgren); Bayer (Andrea Derix, Gunnar Brobert, and Lothar Roessig); Servier (Benoit Tyl); The Hyve (Kees van Bochove); Vifor Pharma (Maureen Cronin). Funders: European Union and the European Federation of Pharmaceutical Industries and Associations Innovative Medicines Initiative, Belgium (Colm Carroll); European Commission, Belgium (Ceri Thompson). Patient Advocacy: European Heart Network, Belgium (Birgit Beger, Susanne Løgstrup, and Marilena Vrana); European Society of Cardiology Patient Forum, France

(Paul McGreavy). Clinical Academic and Professional Societies: European Society of Cardiology, France (Barbara Casadei, Stephan Achenbach, and Valentina Tursini); European Heart Agency, Belgium (Panos E Vardas); Oslo University Hospital and University of Oslo, Norway (Dan Atar); University of Oxford, UK (Colin Baigent); University of Leeds, UK (Chris P Gale); Queen's University Belfast, UK (Donna Fitzsimons); University Hospital Bern, Switzerland (Stephan Windecker); Charité-Universitätsmedizin Berlin, Germany (Stefan D Anker); Royal Brompton Hospital and Imperial College London, UK (Martin Cowie); University College London, UK (Amitava Banerjee, Harry Hemingway, R Tom Lumbers, Spiros Denaxas); University Medical Centre Utrecht, Netherlands (Folkert W Asselbergs, Rick Grobbee, and Ghislaine Van Thiel); University of Birmingham and University Hospitals Birmingham NHS Trust, UK (Dipak Kotecha and George V Gkoutos); University of Cambridge, UK (Angela Wood); Karolinska Institutet and Karolinska University Hospital Stockholm, Sweden (Martin Ingvar). Administrative: BigData@Heart Project Management, Germany (Carl Steinbeisser and Ana Petrova); European Society of Cardiology Project Management, France (Cinzia Ceccarelli, Katija Baljevic, and Polyxeni Vairami); Medical Writer, UK (Jennifer Taylor).

References

1 Conrad N, Judge A, Tran J, et al. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *Lancet* 2018; **391:** 572–80.

2 Hernandez B, Reilly RB, Kenny RA. Investigation of multimorbidity and prevalent disease combinations in older Irish adults using network analysis and association rules. *Sci Rep* 2019; **9:** 14567.

3 Oni T, Youngblood E, Boulle A, McGrath N, Wilkinson RJ, Levitt NS. Patterns of HIV, TB, and non-communicable disease multi-morbidity in peri-urban South Africa—a cross sectional study. *BMC Infect Dis* 2015; **15:** 20.

4 Studer R, Sartini C, Suzart-Woischnik K, et al. Identification and mapping real-world data sources for heart failure, acute coronary syndrome, and atrial fibrillation. *Cardiology* 2022; **147:** 98–106.

5 Wallentin L, Gale CP, Maggioni A, Bardinet I, Casadei B. EuroHeart: European unified registries on heart care evaluation and randomized trials. *Eur Heart J* 2019; **40:** 2745–49.

6 Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence—what is it and what can it tell us? *N Engl J Med* 2016; **375:** 2293–97.

7 John W, Ian H, Rafael J, et al. Implementation and relevance of FAIR data principles in biopharmaceutical research and development. *Drug Discov Today* 2019; **24:** 933–38.

8 von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007; **370:** 1453–57.

9 Nicholls SG, Quach P, von Elm E, et al. The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) statement: methods for arriving at consensus and developing reporting guidelines. *PLoS One* 2015; **10:** e0125620.

10 Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* 2020; **370:** m3164.

11 Logullo P, MacCarthy A, Kirtley S, Collins GS. Reporting guideline checklists are not quality evaluation forms: they are guidance for writing. *Health Sci Rep* 2020; **3:** e165.

12 Dewey M, Levine D, Bossuyt PM, Kressel HY. Impact and perceived value of journal reporting guidelines among radiology authors and reviewers. *Eur Radiol* 2019; **29:** 3986–95.

13 Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016; **3:** 160018.

14 Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health* 2014; **11:** 5170–207.

15 Peng M, Lee S, D'Souza AG, Doktorchik CTA, Quan H. Development and validation of data quality rules in administrative health data using association rule mining. *BMC Med Inform Decis Mak* 2020; **20:** 75.

16 Harron K, Dibben C, Boyd J, et al. Challenges in administrative data linkage for research. *Big Data Soc* 2017; **4:** 2053951717745678.

17 Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care data ran into trouble. *J Med Ethics* 2015; **41:** 404–09.

18 Muller SHA, Kalkman S, van Thiel GJMW, Mostert M, van Delden JJM. The social licence for data-intensive health research: towards co-creation, public value and trust. *BMC Med Ethics* 2021; **22:** 110.

19 Kotecha D, Bax JJ, Carrera C, et al. Roadmap for cardiovascular education across the European Society of Cardiology: inspiring better knowledge and skills, now and for the future. *Eur Heart J* 2019; **40:** 1728–38.

20 Treweek S, Bevan S, Bower P, et al. Trial forge guidance 1: what is a Study Within A Trial (SWAT)? *Trials* 2018; **19:** 139.

21 Wang X, Mobley AR, Tica O, et al. Systematic approach to outcome assessment from coded electronic healthcare records in the DaRe2THINK NHS-embedded randomised trial. *medRxiv* 2022; published online May 25. https://doi.org/10.1101/2022.05.24.22275434 (preprint).

22 Aitken M, Tully MP, Porteous C, et al. Consensus statement on public involvement and engagement with data-intensive health research. *Int J Popul Data Sci* 2019; **4:** 586.

23 Bunting KV, Stanbury M, Tica O, Kotecha D. Transforming clinical research by involving and empowering patients—the RATE-AF randomized trial. *Eur Heart J* 2021; **42:** 2411–14.

24 Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. *J Intern Med* 2013; **274:** 547–60.

25 European Patient Innovation Summit. Position paper on digital health from a patient perspective. 2019. https://www.episummit.net/position-paper.html (accessed Aug 11, 2022).

26 Springate DA, Kontopantelis E, Ashcroft DM, et al. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PLoS One* 2014; **9:** e99825.

27 Bohensky MA, Jolley D, Sundararajan V, et al. Data linkage: a powerful research tool with potential problems. *BMC Health Serv Res* 2010; **10:** 346.

28 Frobert O, Lagerqvist B, Olivecrona GK, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. *N Engl J Med* 2013; **369:** 1587–97.

29 Karwath A, Bunting KV, Gill SK, et al. Redefining β-blocker response in heart failure patients with sinus rhythm and atrial fibrillation: a machine learning cluster analysis. *Lancet* 2021; **398:** 1427–35.

30 WHO. Ethics and governance of artificial intelligence for health: WHO guidance. 2021. https://www.who.int/publications/i/item/9789240029200 (accessed Aug 11, 2022).

31 Wood A, Denholm R, Hollings S, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021; **373:** n826.

32 Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; **340:** c332.

33 von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007; **147:** 573–77.

34 Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015; **12:** e1001885.

35 Chan AW, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Ann Intern Med* 2013; **158:** 200–07.

36 Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med* 2010; **8:** 24.