ELSEVIER

CrossMark

Journal of Clinical Epidemiology 91 (2017) 13–22

Journal of Clinical Epidemiology

Series: Pragmatic trials and real world evidence: Paper 8. Data collection and management

Anna-Katharina Meinecke[a,*], Paco Welsing[b], George Kafatos[c], Des Burke[d], Sven Trelle[e], Maria Kubin[f], Gaelle Nachbaur[g], Matthias Egger[h], Mira Zuidgeest[b], on behalf of work package 3 of the GetReal consortium

[a]*Bayer AG, Global RLE Strategies & Outcomes Data Generation, Aprather Weg 18a, 42096 Wuppertal, Germany*
[b]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, Utrecht, 3508 GA, the Netherlands*
[c]*Amgen Ltd, Centre for Observational Research, 1 Uxbridge Business Park, Sanderson Road, Uxbridge UB8 1DH, UK*
[d]*GlaxoSmithKline PLC, Clinical, Medical and Regulatory IT, R&D IT, Priory Street, Ware, Hertfordshire SG12 0DP, UK*
[e]*Clinical Trial Unit Bern, University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland*
[f]*Bayer AG, Market Access, Aprather Weg 18a, 42096 Wuppertal, Germany*
[g]*GlaxoSmithKline France, PharmacoEPIdémiologie et Modélisations Médico-Economiques, 23, rue François Jacob, 92500 Rueil-Malmaison, France*
[h]*Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland*

Accepted 10 July 2017; Published online 14 July 2017

Abstract

Pragmatic trials can improve our understanding of how treatments will perform in routine practice. In a series of eight papers, the GetReal Consortium has evaluated the challenges in designing and conducting pragmatic trials and their specific methodological, operational, regulatory, and ethical implications. The present final paper of the series discusses the operational and methodological challenges of data collection in pragmatic trials. A more pragmatic data collection needs to balance the delivery of highly accurate and complete data with minimizing the level of interference that data entry and verification induce with clinical practice. Furthermore, it should allow for the involvement of a representative sample of practices, physicians, and patients who prescribe/receive treatment in routine care. This paper discusses challenges that are related to the different methods of data collection and presents potential solutions where possible. No one-size-fits-all recommendation can be given for the collection of data in pragmatic trials, although in general the application of existing routinely used data-collection systems and processes seems to best suit the pragmatic approach. However, data access and privacy, the time points of data collection, the level of detail in the data, and the lack of a clear understanding of the data-collection process were identified as main challenges for the usage of routinely collected data in pragmatic trials. A first step should be to determine to what extent existing health care databases provide the necessary study data and can accommodate data collection and management. When more elaborate or detailed data collection or more structured follow-up is required, data collection in a pragmatic trial will have to be tailor-made, often using a hybrid approach using a dedicated electronic case report form (eCRF). In this case, the eCRF should be kept as simple as possible to reduce the burden for practitioners and minimize influence on routine clinical practice.

Keywords: Pragmatic trial; Routinely collected data; Electronic health records; Registries; Claims databases; eCRF

1. Introduction

Clinical trials that evaluate new drug treatments are required to underpin market authorization and are generally conducted under highly controlled circumstances. Such traditional, explanatory trials deliver data on efficacy and safety of treatments, yet they often insufficiently inform physicians, policy makers, and other stakeholders how treatments will actually perform in real-world clinical practice [1]. Pragmatic trials, however, evaluate relative effectiveness under conditions routinely encountered in clinical practice and can yield evidence of the added value of new treatments in real life [2].

Clinical trials commonly require the collection of highly detailed patient data. This necessitates onsite staff training regarding data collection and management, which presents a burden to study sites, interferes with usual care, and often

Funding: The work leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115546, resources of which are composed of financial contributions from the European Union's Seventh Framework Programme (FP7/2007/2013) and EFPIA companies' in-kind contribution.

* Corresponding author. Tel.: +49 202 36 5455; +49 175 310 9182.
E-mail address: anna-katharina.meinecke@bayer.com (A.-K. Meinecke).

http://dx.doi.org/10.1016/j.jclinepi.2017.07.003

What is new?

Key findings

- There is a need for more pragmatically designed clinical trials to evaluate real-life treatment effects in routine clinical practice in an earlier phase in the drug development process.
- In general, the use of electronic case report forms (eCRFs) for data collection, in the way they are currently being used in most clinical trials, interferes with routine clinical practice. This may lead to the inclusion of a nonrepresentative sample of patients, physicians, and practices and therefore affect the generalizability of the trial results. As such their suitability for pragmatic trials is questionable.

What this adds to what was known?

- This paper addresses challenges and solutions for data collection and management in pragmatic trials, where a high level of accuracy and completeness of data needs to be balanced with a low level of interference with clinical practice.
- Data access and privacy, the time points of data collection, the level of detail in the data, and the lack of a clear understanding of the data-collection process were identified as main challenges for the usage of routinely collected data, for example, electronic health records, in pragmatic trials.
- Current international initiatives, that might facilitate a more pragmatic approach for data collection and management, are discussed.

What is the implication and what should change now?

- The quality of routinely collected data can be optimized by, for example, automated query generation and pop-ups that are embedded in the electronic health record or other electronic health care systems.
- Pragmatic trials should interfere with routine care as little as possible while ensuring high-quality data. Successful existing pragmatic trials typically use a hybrid approach for data collection: dedicated eCRFs combined with routinely collected data.

excludes research-naïve practitioners and sites from participation in such trials [3,4]. This does not affect the aim of more explanatory trials, where the interest is in the pharmacological effect of the drug. However, it may limit the generalizability of the study results and thus the value of evidence of a pragmatic trial. Therefore, pragmatic trials need to closely align all study-related interventions with routine clinical practice, to minimize interference with routine care and enhance the generalizability of the results to the real-word setting [1], while simultaneously ensuring high-quality data.

This paper identifies existing methods for the collection of data and their suitability for pragmatic trials, discusses their respective methodological and operational challenges, suggests possible solutions, and identifies opportunities and developments in this field (Box 1).

2. Methodological choices on data collection for pragmatic trials

The decision on the most suitable method for data collection in pragmatic trials should be based on several considerations, of which the most important ones are described in the following sections.

2.1. Data-collection methods

In general, three approaches for data collection in trials can be identified:

i. Collecting all necessary data using electronic case report forms (eCRFs) that are specifically created for the study. This is the usual method for traditional more explanatory trials. This approach is the most likely to interfere with routine practice, as additional labor and expertise is often required to implement the data-collection process.
ii. Extracting routinely collected data from existing sources, such as electronic health records (EHRs: systems into which practitioners enter routine clinical and laboratory data during usual practice), registries (any system that collects uniform data), or insurance claims (databases maintained by payers for reimbursement purposes or other routine health care databases). This approach does not interfere with routine practice as it does not require additional data collection.
iii. Combining routinely collected data with additional data collected specifically for the study in a hybrid approach.

Which data are required, the level of detail needed, as well as the frequency and timing of collection drive the choice for the method of data collection. Although large numbers of variables are usually collected at predefined time points in an explanatory trial, the data collection in a pragmatic trial would ideally follow real-life clinical practice: that is, data, with the level of detail needed for treating the patient, are collected when patients have contact with their health care provider, who then routinely enters it into his/her health-record system. To find the right balance between the highly controlled data collection in

traditional more explanatory trials and routinely executed data collection (eg, EHRs) remains a challenge when designing a pragmatic trial (Fig. 1). Using a hybrid approach, as generally used in existing pragmatic trials, where "traditional" methods (ie, eCRFs) are used in a limited manner and combined with direct use of routinely collected data may provide a good balance between data quality and interference with routine clinical practice (Boxes 2 and 3) [5,6].

2.2. Quality of data collection

Other aspects of data collection that need to be considered and decided on when designing a pragmatic trial are (1) the extent to which onsite staff will be trained in data collection, (2) the extent of quality checks implemented in the entry system, both directly and remotely, (3) other measures to minimize data-entry errors. The advantages in terms of improved data quality need to be weighed against the possible disadvantage of reducing generalizability because of changed routine practice. Of note, regulatory authorities and ethics committees may demand training on data collection and methods for quality checking [7,8].

3. Operational challenges of data collection and possible solutions

A number of operational challenges can be identified for the different approaches for data collection in more pragmatic trials.

3.1. Challenges with eCRF-based data collection

eCRFs are the most-used tool in more traditional explanatory trials. Their great advantage is that they allow the definition and collection of the exact data required at distinct times (Table 1). Because of the fact that the data are specifically collected for the research purpose, usually collected by trained staff, and typically monitored and validated closely, there is a general perception that this method delivers high-quality, valid data. Nevertheless, data collected by eCRFs also have quality issues, such as entry errors or transcription errors from paper source [9]. Moreover, this method of data collection typically cannot be implemented without at least some interference in routine care. Because of an increase in workload that is generally associated with this approach, the conditions necessary for using eCRFs may often only be fulfilled at specialized sites that have the expertise and manpower to meet all requirements. When detailed data collection is required on top of the usual workload, general practitioners or smaller hospitals, which often treat the patient population of interest, may not be able to cope with this approach [5,10]. In addition, (extensive) staff training in the use of data-collection systems as well as changes in data handling as such may influence routine clinical practice due to the Hawthorne effect [3]. Both might limit the generalizability of the results of a trial. To reduce the burden for practitioners and to interfere with routine clinical practice as little as possible, in pragmatic trials, the eCRF and its implementation should be kept as simple as possible [11]. Moreover, any additional interference, such as onsite staff training and different approaches to minimize and control for data-entry errors, should be kept as small as possible. In addition, one should consider to what extent the required information can be obtained from other systems.

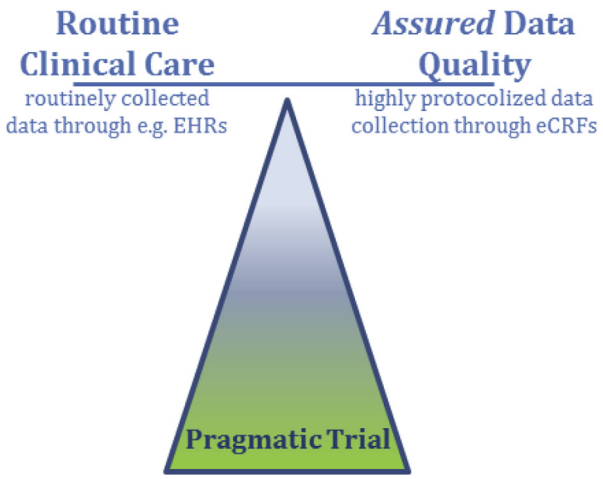

Fig. 1. In a pragmatic trial, the aim is to generate high-quality data and simultaneously interfere with routine care as little as possible. eCRF, electronic case report form; EHR, electronic health record.

Box 1 Series on pragmatic trials

Pragmatic trials aim to generate real-world evidence on the (relative) effects of treatments, generalizable to routine practice. In this series, we will discuss options and choices for pragmatic trial design, operational consequences, and the interpretation of results.

1. Introduction
2. Setting, sites, and investigator selection
3. Patient selection challenges and consequences
4. Informed consent
5. Usual care and real life comparators
6. Outcome measures in the real world
7. Safety, quality and monitoring
8. **Data collection and management**

3.2. Challenges with routinely collected data

The advantage of using routinely collected data in a pragmatic trial is that data collection does not or only minimally interfere with usual care (Table 1). Another advantage of standard electronic systems, especially EHRs, might be that they allow access to numerous variables that were not included in the trial protocol, which nevertheless

Box 2 eLung/Retropro—pragmatic trials using EHRs in combination with add-on systems

Aim: To develop and evaluate methods to implement simple pragmatic trials using routinely collected electronic health records (EHRs) and recruiting patients at the point of care; to identify the barriers and facilitators for general practitioners (GPs) and patients and the experiences of trial participants.

Outcomes: Successful trial completion with implementation of information technology (IT) system for flagging and data processing and documentation of operational and scientific experiences

Approach for data collection:

- Trials were conducted in English and Scottish general practices that contributed their EHRs to a research database (including a total of 459 practices).
- Additional software was developed that allowed for instantaneous monitoring of EHR activities, flagging and clinician's notification during consultation of trial eligibility, complex eligibility assessments using the EHR database, daily eligibility review, confirmation of eligibility and randomization on the study web site, daily monitoring of side effects, and long-term follow-up of major clinical outcomes.
- Recruitment of patients was done by general practitioners (GPs) via invitations (i.e., cold recruitment) and/or immediate flagging (through the HER system) while unscheduled consultation (i.e., hot recruitment).
- GPs were required to complete Web-based protocol and good clinical practice training and required governance approvals were obtained for both trials.
- Potentially eligible patients were identified through the EHR; however, GPs had to confirm eligibility and patients were then randomized using a concealed allocation schedule.
- Neither patients nor clinicians were blinded.
- Central data monitoring was used, and site visits by research staff were not intended but partially conducted when IT/software issues occurred.

Reference: van Staa et al. The opportunities and challenges of pragmatic point-of-care randomized trials using routinely collected electronic records: evaluations of two exemplar trials. HTA 2014; 18(43).

in hindsight could be regarded as potentially relevant information, for example, in the context of modification of treatment effects. However, using routinely collected data may lead to a number of challenges that need to be considered when designing a pragmatic trial as described in the following [12].

3.2.1. Routinely collected data are not meant for research purposes

Systems that collect patient data in routine daily care include EHRs, insurance claims, as well as other health care or vital statistics databases and registries. EHRs are systems in which practitioners enter routine clinical and laboratory data of their patients over the course of their usual practice. Because the primary focus of such data is the support of clinical care rather than research, data may lack detailed information on indications, patient characteristics, treatments, and events and may be less structured, for example only provided as free text [5,12,13]. Claims databases are run by insurers or other payers for reimbursement purposes and may contain information on the diagnoses, procedures, national drug codes, service provider, prescribing physicians, or health plans. This information is usually entered in a structured, coded format, but it may not be very detailed. Consequently, claims databases often lack information on relevant clinical variables and patient characteristics, and often the information will not be timely [12−14]. There is a trend though among database owners of claims data, especially in the United States, to increase the number of clinical variables to make the databases usable for scientific research purposes (eg, Optum) [15].

3.2.2. Generation of valid, accurate, and complete data may be challenging

A major challenge when using routinely collected data is to produce valid, accurate results [12,16]. As data collection happens under real-life conditions, measurement may be more variable and higher levels of missing data and entry errors may be expected, possibly resulting in bias especially, but not only, when measurement error is not random [17,18]. Such bias can call into question the validity of a trial, especially if this ascertainment differs across comparison groups, for example, when physicians suspect an effect on blood pressure of a new treatment, data on blood pressure may be collected more frequent and uniform for that trial arm. In case data are likely to be missing at random, the problem may be overcome by appropriate statistical methods such as multiple imputation [19]. Random errors in data collection and missing data can reduce the power of the study, with implications for the calculation of the required sample size. Gauging the pros and cons of using statistical approaches to deal with the limitations of routinely collected data vs. the implementation of additional data collection is difficult to evaluate

Box 3 Salford Lung Study—a pragmatic trial using EHRs in combination with add-on systems

Aim: The Salford Lung Study (SLS) is the world's first pragmatic RCT (pRCT) of an investigational medication. SLS will evaluate the effectiveness and safety of a new inhaled medicine combination compared with patients' usual care in a broad group of COPD patients in an everyday clinical practice setting. The study is being conducted in and around Salford, UK.

Approach for data collection:

- The SLS uses Salford's electronic health records (EHRs) systems and is delivered in collaboration with local health care providers to allow patients on the study to be closely monitored in real time, but with minimal intrusion into their everyday lives.
- The study's principal investigators are the GPs. They are ideally placed to facilitate recruitment, identify, and report SAEs or serious ADRs and report study end points. GPs may make treatment adjustments according to their clinical opinion. Repeat prescriptions of study medication are issued by GPs as usual and collected by patients from their usual pharmacy.
- Pharmacy data were initially collected by faxing copies of study treatment prescriptions to the study coordination center, but these are now collected electronically. Prescription collection data are used to assess treatment adherence.
- Hospital admissions are primarily to two local hospitals, Salford Royal Hospital and the University Hospital of South Manchester. Relevant admissions are identified electronically and assessed by a separate safety team.
- The integrated system was set up by NorthWest EHealth, a not-for-profit organization formed by a partnership between The University of Manchester, Salford Royal Foundation Trust and Salford Clinical Commissioning Group, and links patients' records across their GP surgeries and hospitals. This removes the need for the enforced interventions and controls required in RCTs which may affect the way a patient behaves—for example, how and when they take their medication.

References: The Salford Lung Study protocol: a pragmatic, randomized phase III real-world effectiveness trial in chronic obstructive pulmonary disease.

Nawar Diar Bakerly, Ashley Woodcock, John P. New, J. Martin Gibson, Wei Wu, David Leather and Jørgen Vestbo. Respiratory Research 2015;16:101 http://dx.doi.org/10.1186/s12931-015-0267-6.

Clinical trials meet the real world—https://www.gsk.com/en-gb/behind-the-science/patients-and-consumers/clinical-trials-meet-the-real-world/

Box 4 TASTE—a pragmatic trial nested in an existing registry

Aim: To evaluate the effects of thrombus aspiration in patients with ST-elevation myocardial infarction undergoing a primary percutaneous coronary intervention

Outcomes: All-cause mortality (primary), coronary re-interventions, and outcomes related to the intervention

Approach for data collection:

- The trial was embedded in the Swedish Coronary Angiography and Angioplasty Registry (SCAAR).
  a. the registry is fully embedded in daily clinical practice with case report forms implemented electronically and Web-based data entry
  b. included all 29 Swedish centers performing coronary angiography and interventions
  c. collects detailed information about coronary interventions over time
  d. linked to the Swedish National Population Registry including personal data and deaths
  e. regular source data verification on randomly sampled centers/variables.
- Randomization for the trial was directly implemented in the eCRFs of the registry, and all outcomes of the trial were based on data routinely collected within the registry
- Due to this approach, the trial could be performed without disruption of usual practice, only informed consent and randomization deviated from routine practice.

Reference: Frobert O. et al. Thrombus aspiration during ST-segment elevation myocardial infarction. N Engl J Med 2013; 369: 1587-97.

Box 5 SCOT—a pragmatic trial using record linkage for identifying potential outcome events

Aim: To evaluate the cardiovascular (primary) and gastrointestinal (secondary) safety of celecoxib vs. traditional nonsteroidal anti-inflammatory drug therapy in patients with osteoarthritis or rheumatoid arthritis.

Outcomes: First hospitalization for nonfatal myocardial infarction or nonfatal stroke or cardiovascular death. Other outcomes were hospitalization or death for various reasons (gastrointestinal, specific cardiovascular, and renal)

Approach for data collection:

- Participants were recruited in general practices (Scotland, England, Denmark, and Netherlands) and randomized by telephone or a web site.
- Outcome events were not directly collected during follow-up but potential events identified through record linkage: the trial database was linked every 3 months to national population health care and mortality databases containing all hospital discharge diagnoses or causes of death in a particular country.
- For each potential outcome event, original case records were requested. Each potential event was adjudicated by a blinded, independent adjudication committee.
- Serious adverse events were recorded in the electronic case report form.

Reference: MacDonald TM et al. Methodology of a large prospective, randomized, open, blinded end point streamlined safety study of celecoxib vs. traditional nonsteroidal anti-inflammatory drugs in patients with osteoarthritis or rheumatoid arthritis: protocol of the standard care vs. celecoxib outcome trial (SCOT). BMJ open 2013; 3: e002295.

beforehand [20]. We recognize that in practice, it may not be easy to assess the influence of errors a priori, and this may need a separate feasibility study. The evaluation of measurement of outcomes for pragmatic trials as performed in practice and measures on how to improve this, both important components of feasibility studies, are addressed elsewhere in this series [21].

Training staff in the use of a routine data-collection system might solve or alleviate the problem of incorrect or incomplete data entry. However, training staff may change more than data entry alone: training causes interference, which might have an impact on behavior in general (Hawthorne effect) [3]. Furthermore, it might lead to changes in treatment, if training relates to information relevant for treatment decisions and thus interfere with the usual care provided to patients, possibly resulting in a decrease in the generalizability of trial results [1].

Another option that might interfere less with clinical routine would be to implement data quality checks in the system that would detect incorrect or missing data during data entry, and specify procedures for correction. When a health care database is the source of study data, any change made to the database after data extraction needs to be captured in an audit trail [22,23].

3.2.3. Quality and completeness of data varies within and among databases

The collected data as such and their level of detail, completeness, and correctness can vary, both within and between existing databases [13,16].

Generally, ‘‘completeness’’ of data in health care databases depends on what data exactly are routinely collected in health records and the exact data needs for a trial, and one should be explicit about both when establishing the appropriateness of health care databases as a data source for a pragmatic trial [24]. In addition, before selecting a data-collection approach solely based on routinely collected data, one needs a very good understanding of the process of data entry and management in the individual database [25] as well as of the database landscape within and across settings and countries in which a pragmatic trial will be performed [26]. As a starting point for finding the optimal database for ones demands, the International Society for Pharmacoeconomics and Outcomes Research provides an international digest of databases [27]. This list is neither exhaustive nor does it assess the quality, completeness, and accessibility of data and its potential for customization for a specific study, which would assist the assessment of the suitability of data for inclusion in pragmatic trials. However, guidelines exist for how to select a database for a study, which are yet oriented toward noninterventional studies [28]. The Medicines & Healthcare products Regulatory Agency has recently published a position paper in which compliance issues of EHRs are identified and user requirements for EHR owners are formulated to ensure good clinical practice compliance [29]. It provides a number of explicit recommendations, such as ‘‘Audit trails for information added to the EHR. Any new information added to the subjects’ medical notes (whether paper or electronic) should show when the entry was made and by whom, so that the documentation provides a full audit trial of events (any amendments/deletions etc.).’’ This paper might be used as a basis for quality checks of EHRs.

3.2.4. Compliance with safety requirements on adverse event reporting

As described previously, routine data are typically collected infrequently or irregularly, which poses problems

Table 1. Approaches for data collection for pragmatic trials with their potential (dis)advantages

Considerations for data collection	Use of eCRF solely	Extraction of routinely collected data	Use of hybrid approach
Interference with routine clinical practice → limits generalizability of results	May exclude sites from participating due to extra workload and change routine clinical practice, depending on how comprehensive the additional workflow will be.	Does not change routine clinical practice	Depends on i. how extensive the additional data collection is ii. how well data validation checks are already implemented in existing system iii. how comprehensive the additional workflow will be
Time points of data collection	Usually optimized to meet study needs but could also be aligned with routine practice	Dictated by routine clinical practice	Can be optimized to meet study needs but could also be aligned with routine practice
Level of detail and quality of the data	Level of detail is optimized for the study. In general, high-quality valid data, but entry/transcription errors, may occur	Data are not collected for research purposes and may be more variable, display more missing data and entry errors (both random and systematic) may be expected. Level of detail might be insufficient for trial	In between eCRF and the approach of using routinely collected data depending on specific approach
All relevant variables collected	Yes as eCRF is made specifically for the study	Possibly not. However, one advantage is that variables that are not collected as part of the trial but are relevant in hindsight might be present in the database	Probably yes as the hybrid approach is designed to obtain sufficient quality data on all relevant variables for the study, making use of routinely collected data as much as possible.
Compliance with safety requirements—reporting of adverse event	Yes	Possibly not i. data are collected only at routine contact moments between patient and care provider, which may occur infrequent to collect sufficient data on safety ii. data extraction may often not be in real time, as such mandatory timelines for safety reporting may not be met	Probably yes: safety requirements may necessitate i. the implementation of extra data collection when frequency of patient visits is too low or if data collected are not sufficient ii. an adjustment in the frequency of data extraction or upload from EHR if the routine system is not sufficient
Data privacy and access	Not an issue—only study-specific data are collected in line with informed consent for the study.	Access to routinely collected data(base) gives access to data beyond what is required for a study which introduces an ethical challenge.	Access to routinely collected data(base) gives access to data beyond what is required for a study which introduces an ethical challenge.

Abbreviations: eCRF, electronic case report form; EHR, electronic health record.

particularly for the collection of adverse events (AEs), for which frequent or regular scheduled visits may be deemed necessary. Especially delays in recording data in relevant databases and time between data entry and information becoming available to the sponsor/in the study database can interfere with obligatory timelines for reporting AEs as well as result in missing data. This can adversely influence the validity of the study and raise safety issues [7,30]. Calling patients after a defined time frame without GP interaction might be a valid approach of ensuring the assessment of AEs while keeping interference limited [6].

3.2.5. Data privacy and access

A further challenge with using routinely collected data is protection of the privacy and confidentiality of research participants [12]. Providing access to a health care database system and therefore to patient data beyond what is required for a study, as opposed to recording and accessing study-specific data in an eCRF raises right-to-privacy concerns [31]. In a recent series of papers exploring ethical issues in pragmatic trials, McGraw et al. point out that patients have concerns about privacy protection and may engage in "privacy-protecting behavior" such as withholding information to health care providers. The authors argue that neither privacy nor the value of pragmatic trials should be "compromised for the sake of the other" [32]. Traditionally, anonymity and informed consent have protected patient information, but it has been argued that neither can ensure privacy in the era of big data [32,33]. Technological solutions are sought to fill the protection gap, but it is unlikely that these solutions will defeat all challenges. To allow the use of personal data, modifications of informed consent procedures have been

proposed [34]. The concept of a learning health care system also has emerged in which knowledge generation is "a natural outgrowth and product of the health care delivery process" [35]. In such a system, accessing patient data for research purposes is likely to be facilitated. However, because a tradeoff between utility and security remains and guidance on how to balance privacy and research interests is currently lacking, a small risk of identification might remain [36].

Privacy concerns inevitably constrain access to data by investigators. Currently, access is often granted to database owners only, decreasing the possibility of using routinely collected data. In some countries, national, local, or regional data privacy regulations prevent the use of routinely collected patient-level data and the linkage of information on individuals between databases such as death or disease registries. Apart from the data privacy issue, technical challenges exist, for example, efficient and timely data extraction from systems can be particularly challenging when data must be de-identified but the possibility of individual patient identification needs to be retained for safety reasons. In some settings, algorithms for probabilistic linking have been developed to address this [37].

4. Opportunities and current developments for the use of health care databases for pragmatic trials

4.1. International initiatives support the utilization of health care databases

Several initiatives worldwide are dedicated to improving both the standardization and quality of routinely collected data. As routinely collected data were not intended for research purposes, it is important to create a joint understanding on the do's and don'ts of its collection and use. Therefore, several guidance documents have been or will be published to ensure standards and best practices when using routinely collected data [29,38]. In addition, several initiatives work on technical solutions for challenges that arise when using EHRs or other health care databases for research. Supported by the EU's 7th Framework Programme, the TRANSFoRm project aims to develop a "rapid learning healthcare system" that can improve both patient safety and the conduct and volume of clinical research in Europe. A dynamic interface is integrated with EHRs to identify patients eligible for research and collect outcomes and safety data. Improving the interoperability of EHR data and other data sources is another aim of TRANSFoRm [39]. The Electronic Health Records for Clinical Research (EHR4CR) project of the EU's Innovative Medicines Initiative developed a technological platform that combines hospital data across countries, with a focus on the identification of sites and patients for trials [40]. Another approach to tackle the challenge of heterogeneity of multiple sources is being developed in the United States under the FDA's Sentinel Initiative, which aimed to develop a linked system that will draw on existing health care data from multiple database sources to actively monitor the safety of medical products in real time [41]. PCORnet again creates clinical data and patient-powered research networks to further enhance effectiveness research [42]. To enhance the quality of EHRs and make their level of quality more visible, the eClinical consortium will release a tool under the EHRCR initiative, which will allow doctors to evaluate the quality and security of their own EHR systems and provide study teams information on the quality and content of EHR systems of interest [43].

Achieving homogeneous quality of databases on an international scale will always be challenging and a formal check of data quality and completeness will be needed to examine the suitability of EHR systems or sites for pragmatic trials.

4.2. The use of existing registries

A special source of routinely collected data is existing registries [44]. For the purposes of this article, we define a registry as any system that collects uniform data [45]. Registries are becoming increasingly common, aiming at high if not 100% coverage of patients with a target disease or treatment. If high coverage is achieved, registries may be considered (public health) surveillance systems. Participation in a registry is often embedded within clinical practice either via an additional eCRF or by use of a standard IT system. If a registry for a specific patient population of interest exists, it may be particularly suited to a "nested" pragmatic trial or so-called "registry-based" randomized trial [46–48]. For example, within the TASTE trial, which is nested within the Swedish Coronary Angiography and Angioplasty Registry, one out of several registries that run under the SWEDEHEART registry, all data on end points were collected through the "routine registry data collection" (Box 4) [47]. This did not just ease patient recruitment and data collection to a great extent but ultimately reduced the costs per patient tremendously (50$ per patient) [49]. Registries may be open to adapting data collection if relevant for the research objective, especially if also relevant for patient care. Given that data-collection processes are already in place within registries, implementing changes may be easier than setting up a data-collection framework from scratch (eg, implementing eCRFs). The European PARENT initiative (cross-border PAtient REgistries iNiTiative) aims to support the development of comparable and interoperable patient and disease registries in EU Member States. PARENT is creating a "registry of registries," which may soon provide an overview of the European registry landscape, facilitating the identification of relevant registries [50].

4.3. The use of a hybrid approach

"Hybrid approach" means that routinely collected data are complemented with additional data collected for a specific study, either through the implementation of additional IT-solutions to the existing EHR, linking multiple

databases, collecting patient reported outcomes [51], or by implementing additional eCRFs [6]. These may be solutions to the challenges that arise when using either eCRFs or existing medical records solely for data collection, including the limited availability of certain variables, insufficient detail, and the timing of data collection but also an increased interference with routine clinical practice. A hybrid approach may enable an appropriate data-collection framework which takes advantage of both worlds while reducing overall workload. The eLung and Retropro study are two examples where the implementation of IT systems within existing EHRs made it possible to flag eligible patients, efficiently evaluate eligibility, randomize patients, collect outcome data, and monitor adverse events (Box 2) [5]. In contrast, the Salford Lung study used near real-time linkage of primary as well as secondary care electronic health records next to limited eCRFs for data collection (Box 3) [6]. The SCOT trial used linkage to national population health care and mortality databases to ascertain outcome regarding hospitalizations and a separate eCRF for adverse events (Box 5) [52].

To integrate additional data, eCRFs or add-on solutions into an existing system, access to the back-end data of and IT information pertaining to an EHR or health care database is a prerequisite. The integration of add-on systems may make additional data entry easier, compared to using extra eCRFs, and some quality control, for example, using reminders, might be added while still appropriately reflecting usual care. This could be done, for example, via pop-up windows, which would have to be completed to be closed, requesting additional information needed for the trial or reminding the GP to schedule follow-up visits [5].

5. Conclusions and recommendations

Decisions regarding data collection and management are important in any clinical trial. When designing pragmatic trials, it needs to be decided on how much interference with routine clinical practice through the data-collection process one is willing to accept to obtain high-quality data. Such interference should be as limited as possible and as such, the use of existing EHRs or other routinely used electronic systems for data collection and management should always be considered. To determine the feasibility of using standard systems, one needs to evaluate the process of data entry and the level of detail and completeness of the data of these systems to compare the quality of their data with that needed for the trial. If an (additional) eCRF-based data-collection process is chosen, its structure and implementation should be as easy and as much in line with routine clinical practice as possible. The challenges and possible solutions identified in this paper should help design teams to assess both opportunities and limitations in the long list of potential tools for data collection and management in their pragmatic trials and to make evidence-based decisions about the most appropriate method for data collection.

Acknowledgments

The research leading to these results was conducted as part of the GetReal consortium and included literature review and interviews with stakeholders from academia, research institutions, contract research organizations, the pharmaceutical industry, regulatory authorities, health care insurers, health technology assessment (HTA) agencies, general practitioners, and patient organizations. For further information, please refer to http://www.imi-getreal.eu/.

References

[1] Zuidgeest MGP, Goetz I, Groenwold RHH, Irving E, van Thiel GJMW, Grobbee DE. Series: Pragmatic trials and real world evidence: Paper 1. Introduction. J Clin Epidemiol 2017;88:7−13.
[2] Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis 1967;20:637−48.
[3] McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. J Clin Epidemiol 2014;67:267−77.
[4] Martin K, Bégaud B, Latry P, Miremont-Salamé G, Fourrier A, Moore N. Differences between clinical trials and postmarketing use. Br J Clin Pharmacol 2004;57:86−92.
[5] van Staa TP, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B, et al. The opportunities and challenges of pragmatic point-of-care randomized trials using routinely collected electronic records: evaluations of two exemplar trials. Health Technol Assess 2014;18:1−146.
[6] New JP, Bakerly ND, Leather D, Woodcock A. Obtaining real-world evidence: the Salford Lung Study. Thorax 2015;70:1008.
[7] Irving E, van den Bor R, Welsing P, Walsh V, Alfonso-Cristancho R, Harvey C, et al. Series: Pragmatic trials and real world evidence: Paper 7. Safety, quality and monitoring. J Clin Epidemiol 2017;91:6−12.
[8] ICH harmonised tripartite guideline: guideline for good clinical practice E6/International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH E6, 1 May 1996.
[9] Mitchel JT, Joon Kim Y, Choi J, Park G, Cappi S, Horn D, et al. Evaluation of data entry errors and data changes to an electronic data capture clinical trial database. Drug Inf J 2011;45(4):421−30.
[10] Laken MA, Dawson R, Engelman O, Lovelace O, Way C, Egan BM. Comparative effectiveness in the "real" world: lessons learned in a study of treatment-resistant hypertension. JASH 2013;7(1):95−101.
[11] Raymond J, Darsaut TE, Altman DG. Pragmatic trials can be designed as optimal medical care: principles and methods of care trials. J Clin Epidemiol 2014;67:1150−6.
[12] Coorevits P, Sudgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Int Med 2013;274:547−60.
[13] Simon G, Wagner E, Vonkorff M. Cost-effectiveness comparisons using "real world" randomized trials: the case of new antidepressant drugs. J Clin Epidemiol 1995;48:363−73.
[14] Choudhry NK, Shrank WH. Implementing randomized effectiveness trials in large insurance systems. J Clin Epidemiol 2013;66:S5−11.
[15] Optum. Available at https://www.optum.com/. Accessed July 12, 2017.
[16] Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol 2012;65:343−9. e342.
[17] Buzkova P. Measurement error and outcomes defined by exceeding a threshold: biased findings in comparative effectiveness trials. Pharm Stat 2012;11:429−41.
[18] Lachin JM. The role of measurement reliability in clinical trials. Clin Trials 2004;1:553−66.
[19] van Buuren S. Flexible imputation of missing data. Boca Raton, FL: Chapman & Hall/CRC; 2012.

[20] Janssen KJ, Donders AR, Harrell FE Jr, Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 2010;63:721−7.
[21] Welsing P, Oude Rengerink K, Collier S, Eckert L, van Smeden M, Ciaglia A, et al. Series: Pragmatic trials and real world evidence: Paper 6. Outcome measures in the real world. J Clin Epidemiol 2017; 90:99−107.
[22] FDA, CDER, CBER, CDRH, CFSAN, CVM, ORA. Guidance for industry part 11: electronic records, electronic signatures − scope and application 2003. Available at http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm125125.pdf.
[23] European Commission EudraLex. Volume 4 good manufacturing practice (GMP) guidelines. Annex 11: Computerised Systems 2011. Available at http://ec.europa.eu/health/files/eudralex/vol-4/pdfs-en/anx11_en.pdf.
[24] Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013;46(5):830−6.
[25] Jansen ACM, van Aalst-Cohen ES, Hutten BA, Büller HR, Kastelein JJP, Prins MH. Guidelines were developed for data collection from medical records for use in retrospective analysis. J Clin Epidemiol 2005;58:269−74.
[26] Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol 2017;9:245−50.
[27] ISPOR Digest of Databases. Available at http://www.ispor.org/digestofintdb/countrylist.aspx.
[28] Hall GC, Sauer B, Bourke A, Brown JS, Reynolds MW, LoCasale R. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiol Drug Saf 2012;21:1−10.
[29] MHRA position statement and guidance electronic health records. MHRA; 2015: Version1.0. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/470228/Electronic_Health_Records_MHRA_Position_Statement.pdf.
[30] Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tuni S, Haynes B, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ 2008;337:a2390.
[31] Melton LJ. The threat to medical-records research. NEJM 1997;337: 1466−9.
[32] McGraw D, Greene SM, Miner CS, Staman KL, Welch MJ, Rubel A. Privacy and confidentiality in pragmatic clinical trials. Clin Trials 2015;12:520−9.
[33] Check Hayden E. Researchers wrestle with a privacy problem. Nature 2015;525:440−2.
[34] Kalkman S, van Thiel GJMW, Zuidgeest MGP, Goetz I, Pfeiffer BM, Grobbbee DE, et al. Series: Pragmatic trials and real world evidence: Paper 4. Informed consent. J Clin Epidemiol 2017;89:181−7.
[35] Olsen L, Aisner D, McGinnis JM, editors. Roundtable on Evidence-Based Medicine. The Learning Healthcare System: Workshop Summary. Washington, DC: The National Academies Press; 2007.
[36] Kotecha JA. Ethics and privacy issues of a practice-based surveillance system. Can J Fam Pract 2011;57:1165−73.
[37] Aplenc R, Fisher BT, Huang YS, Li Y, Alonzo TA, Gerbing RB, et al. Merging of the National Cancer Institute-funded cooperative oncology group data with an administrative data source to develop a more effective platform for clinical trial analysis and comparative effectiveness research: a report from the Children's Oncology Group. Pharmacoepidemiol Drug Saf 2012; 21(Suppl 2):37−43.
[38] FDA Guidance for Industry. Use of electronic health record data in clinical investigation. Draft Guidance 2016. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM501068.pdf.
[39] Transform. Available at http://www.transformproject.eu/.
[40] Electronic Health Records for Clinical Research. IMI. Available at http://www.ehr4cr.eu/.
[41] FDA Sentinel Initiative. Available at http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm2007250.htm.
[42] The national patient-centered clinical research network (PCORnet). Available at http://pcornet.org/about-pcornet/.
[43] eClinical Forum. Available at http://eclinicalforum.org/ehrcrproject/en-gb/home.aspx.
[44] Li G, Sajobi TT, Menon BK, Korngut L, Lowerison M, James M, et al. Registry-based randomized controlled trials—what are the advantages, challenges, and areas for future research? J Clin Epidemiol 2016;80:16−24.
[45] In: Gliklich R, Dreyer N, Leavy M, editors. Registries for Evaluating Patient Outcomes: A User's Guide, 13 2014. 3rd ed. Agency for Healthcare Research and Quality (US); AHRQ Methods for Effective Healthcare, 2014. Report No.: 13(14)-EHC111.
[46] Lauer MS, D'Agostino RB. The randomized registry trial—the next disruptive technology in clinical research? N Engl J Med 2013;369: 1579−81.
[47] Frobert O, Lagerqvist B, Olivecrona GK, Omerovic E, Gudnason T, Maeng M, et al. Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction. N Engl J Med 2013;369:1587−97.
[48] Clinical Trial Transformation Initiative (CTTI). Available at https://www.ctti-clinicaltrials.org/projects/registry-trials.
[49] Wachtell K, Lagerqvist B, Olivercrona GK, James SK, Froeber O. Novel trial designs: lessons learned from Thrombus Aspiration during ST-segment Elevation Myocardial Infarction in Scandinavia (TASTE) trial. Curr Cardiol Rep 2016;18:11.
[50] Parent Joint Action. Registry of registries. Available at http://www.parent-ror.eu/about/#/.
[51] Wu AW, Kharrazi H, Ebony Boulware L, Snyder CF. Measure once, cut twice−adding patient-reported outcome measures to the electronic health record for comparative effectiveness research. J Clin Epidemiol 2013;66:S12−20.
[52] MacDonald TM, Machenzie IS, Wei L, Hawkey CJ, Ford I, SCOT study group collaborators. Methodology of a large prospective, randomised, open, blinded endpoint streamlined safety study of celecoxib versus traditional non-steroidal anti-inflammatory drugs in patients with osteoarthritis or rheumatoid arthritis: protocol of the standard care versus celecoxib outcome trial (SCOT). BMJ open 2013;3:e002295.