English reprint of the original German publication

Use of Electronic Health Data in Clinical Development

From Protocol Design to Patient Identification: A Requirements Specification from the User Perspective

Dr. med. Manfred Stapff

TriNetX, Inc., Cambridge, USA

In clinical research and development, the scientific possibilities for analyzing large volumes of data are still not used to the extent that it is possible in other sectors (e.g. finance, consumer behavior). Health data are often widely distributed and locked in individual databases, standards are highly inconsistent, and data privacy protection complicates data consolidation and data use. This results in complex clinical protocols with often unrealistic selection criteria, and trials are still too often assigned to inappropriate sites. Furthermore, patient recruitment continues to be one of the major problems in the execution of clinical trials [1].

The use of electronic health data (real world data) allows alignment of protocols to actual medical conditions, formulation of realistic inclusion and exclusion criteria and testing their effects on recruitment using real data. In addition, trials can be assigned to sites that have a proven number of patients in their databases, and patients can be identified at the site.

Various providers are players in the field of "big data" and it is not always easy to assess which system is best suited to meet the demands of clinical development. Therefore, a requirements specification is presented in the following.

Challenges of Clinical Development

While sectors such as the telecommunications or music industry have undergone radical changes in the last two decades, the pharmaceutical industry has essentially not changed its product development processes for the last 25 years. In other sectors, the greatest advances were achieved by utilizing the enormous possibilities for collecting and processing electronically available information. Data on consumer behavior (surfing the Internet, online shopping) can be collected across borders and opens many possibilities for analysis, targeted advertising and linking of supply and demand. In contrast to consumer or financial data, health data —if even available electronically—is isolated in individual databases, difficult to link due to differing standards [2], fragmented and subject to stringent requirements regarding data privacy protection and secured access.

This has led to a situation where the enormous possibilities of analysis and use of health data in clinical development have scarcely begun to be utilized. Better use of electronic health data could alleviate or even eliminate problems the pharmaceutical industry has faced for decades.

Protocol Development without Real-World Data

Clinical protocols (trial protocols) are typically based on information from literature, recommendations from external specialists (opinion leaders), internal corporate standards or prior studies (often essential portions carried over via "cut and paste" from Phase II to Phase III despite the change of intention). Only rarely do authors of protocols have the opportunity to base a protocol on the way patients actually present themselves with a specific condition (indication) in real medical care.

AUTHOR

Dr. med. Manfred Stapff

is a physician and clinical pharmacologist with over 20 years of experience in the pharmaceutical industry. After his training in internal medicine, he gained extensive knowledge of pharmaceutical development through diverse positions in clinical research, medical services and regulatory affairs in research-oriented pharmaceutical companies such as Merck & Co., Inc. and Forest/Allergan. Dr. Stapff is currently employed as the Chief Medical Officer of the Health Data Network TriNetX in Cambridge, MA, USA.

His publication list encompasses articles and contributions in the field of cardiovascular circulation, clinical trials and GCP as well as two books on pharmaceutical trials and project management.

Consideration of electronic health records (real world data) would allow protocol authors to take into account frequently occurring concomitant diseases in order to assess typical concomitant medication for possible interactions or to investigate exclusion criteria as to how close or far they are from reality. Instead, protocols are developed based on a hypothetical patient found in the literature and rarely tested against real data before the clinical trial is begun. This leads to disappointing progress in recruitment and to expensive amendments that cause delays [3].

■ Non-Targeted Search for Sites with Patients

Like protocol development, the search for study sites is also based much more on previous experience, opinions or recommendations than on real-life data. About 20 % of initiated trial sites do not enroll a single patient in the clinical trial. A historical rule of thumb states that at best 5–10 % of the number of patients initially stated by an investigator are ultimately enrolled in the trial [4]. In contrast, the comparison of inclusion and exclusion criteria in the protocol with the site's specific patient database may increase this estimate to approximately 50 % [5].

■ Inefficient Patient Identification and Recruitment

Sites perform pre-screening (search for suitable patients without special screening tests) still largely manually. If no centrally generated characterization of the selection criteria is available for searches in the local patient database, this will lead to a decentralized and inefficient duplication of efforts at all sites.

Opportunities

Access to representative health records, e.g. diagnoses, laboratory values, concomitant medication, previous illnesses or genetic information provides a better understanding of the actual status of patient cohorts with a specific condition (trial indication). This leads to a more realistic design of inclusion and exclusion criteria and improves the predictability of later recruitment progress. The selection of the study sites can be limited to those organizations that demonstrably have a sufficient number of patients meeting the eligibility criteria. In chronic diseases, i.e. long-term patients, the conventional manual "chart review" can be replaced by an export from the local database with re-identification at the site.

The opportunities and potential efficiency gains that can be realized using electronic health records are enormous. It is therefore not surprising that different providers are rushing into the market in this sector. However, not everything which appears to be technological progress using large volumes of data (big data) leads to efficiency and quality gains in clinical development. It is therefore important to adhere to a requirements specification to avoid disappointments and bad investments.

Requirements and Wish List

■ Applicable to the Entire Process

Ideally a system with electronic health data should be applicable during the entire trial planning and preparation process (Fig. 1) without needing to switch between systems, databases or providers.

Trial Planning

The preparation and planning process for a clinical trial begins with the initial design of the protocol. Normally, the patient group to be investigated is defined by one or two primary inclusion criteria, mostly aris-

■ Figure 1

Process	Protocol Design; Study Design	→	Site Selection	→	Patient Screening
Problem	• Complex Protocols • Selection criteria often scientifically reasonable but unrealistic		• Availability of patients unclear, often only estimated • Approx. 20% of sites recruit no patients		• Pre-screening potential patients based on selection criteria is often a manual and laborious process
Solution	• Protocol development using real-life data • "Simulation" of recruitment prior to study start		• Targeted site selection based on verified numbers of patients fulfilling eligibility criteria		• Re-identification of patients meeting actual protocol specific eligibility criteria

Use of electronic health data in the planning and preparation process of a clinical trial (Source: All figures were made by the Author/ TriNetX Inc.).

Pharm. Ind. 79, Nr. 2, 204–210 (2017)

ing from the targeted indication and based on the planned product profile. Information on the planned patient cohorts typically comes from literature, earlier studies, own experience or information from external experts (opinion leaders). These sources of information do not always reflect the actual picture of the patient group in the "real" world. Access to electronic health data provides realistic information about the age and sex distribution, concomitant diagnoses, concomitant therapy or laboratory findings for the planned patient cohorts and allows use of this data for the conception of future protocols.

Protocol Testing

As soon as the protocol contains all inclusion and exclusion criteria, it can be tested against the electronic health data for feasibility (Fig. 2). This allows testing (simulation) of the recruitment process and identification of particularly problematic (unrealistic) selection criteria. Some criteria may be reasonable from a scientific perspective but may not reflect the real-life situation. This could cause major obstacles later on. With a critical analysis against real health data, obstacles and potential enrollment hurdles can be identified early on and discussed as part of the internal corporate protocol review [6].

Site Selection

After inclusion and exclusion criteria have been defined, a system based on electronic health data will allow identification of those sites that have a large number of patients meeting the defined eligibility criteria. In addition to the total number of potentially suitable patients it is also important that the most recent trends over time can be represented (e.g. along a quarterly axis) to avoid misdirection due to outdated numbers. Networks that already exist between different providers facilitate this cooperation [7].

Patient Screening

Ideally, the reference to an individual patient should never be entirely lost due to data privacy measures such as anonymization. A re-identification code should remain in the data source. This allows patients who meet the selection criteria to be re-identified at the site (behind the firewall) using a code and in accordance with local ordinances and guidelines (e.g. study-specific directions by a local institutional review board or ethics committee). These patients can then be contacted about potential study participation. The efficiency of screenings could be improved considerably in this way. Patient recruitment systems exist that are compatible with internal hospital databases [8] but cover only this "last step" of the study process.

Contains Sufficient Data

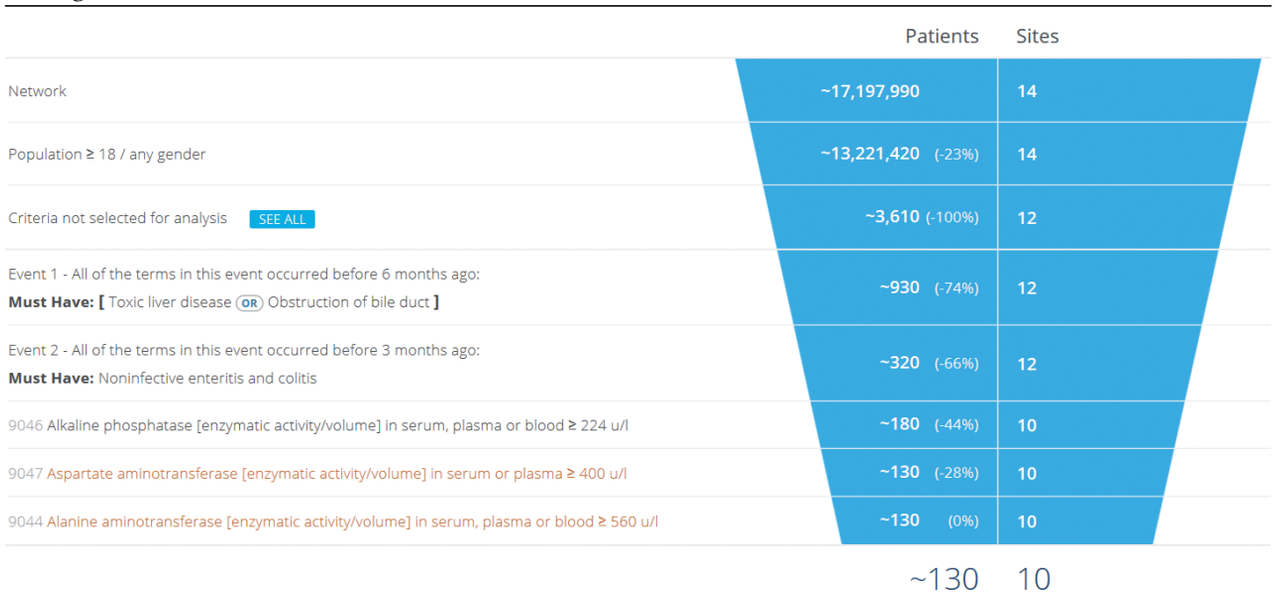
Inclusion and exclusion criteria for clinical trials are predominantly defined using the following data:

- Diagnoses
- Medication
- Lab Results
- History
- Diagnostic Tests
- (Tumor) Mutations

It is self-evident that an electronic health data system is more helpful the larger and the more specific the data volumes it contains [9].

Conditions or test results with large individual variations often need to be rechecked in a screening visit even if historic data are available. Such historic values would better be expressed by their underlying diagnoses or by information about

Figure 2



"Patient Funnel" for testing the effect of selection criteria on patient availability.

corresponding therapy were available. For example, blood pressure values will be re-collected anyway in the patient's untreated condition (e.g., after a washout period), as this usually is an acceptance criterion for clinical trials. Therefore the information on the diagnosis of hypertension (ICD10 I10) is more significant than individual values.

A true goldmine will arise as medical free text fields, such as discharge reports, x-rays or histology findings can be analyzed. Such information is usually made available in form of electronic documents, like PDF files, but in an unstructured way. In the future "Natural Language Processing" (NLP) will be applied so these documents can be scanned, organized in a structured manner, implemented into a database structure, and analyzed.

It is self-evident that the data sources must be regularly updated and originate from a large number of sites, if they are intended for site selection purposes. To ensure data is representative of the entire population and suitable to be used across multiple therapeutic areas, the data volume must exceed a certain critical mass. A number of approx. 7 to 10 million patients is considered a minimum usable size [11].

■ User-Friendliness

Queries in databases require programming efforts. However, the user (scientists, clinicians) cannot always be expected to be versed in XML or SQL code or to constantly have a skilled programmer at hand to generate a quick query. Today, a user interface must be designed for simple operation, so a scientist can use any web browser to access the information from a large data volume without any programming knowledge. As an example, only minimum training should be required to use a graphical user interface for the quick generation of the following lists: all concomitant therapies of a patient cohort with type II diabetes mellitus, elevated creatinine values and hypertension, who have not received Metformin in the last 6 months (this serves only as an example; selection criteria for clinical subjects are obviously much more complex).

■ Minimized Data Security Concerns

Working with and combining millions of medical records (big data) opens up previously unknown possibilities for the identification of diseases, their associated risk factors and causes, including probabilities for recovery. This provides a gateway to personalized medicine that allows targeted treatment based on specific patient characteristics (often of a genetic type). This raises the conflict between fact-based knowledge and the emotion-based push to protect the private sphere. It is not the objective of this article to go into the ethical side of this discussion in detail. However, some facts should be considered in order to minimize potential data privacy concerns when working with electronic health data.

Obviously, the records must be anonymized—cleared of any references to the patient's identity (name, date of birth, address). The data must remain assigned to an individual because it is actually the combination of different information coming from one and the same (unknown, anonymous) patient that allows scientific conclusions to be made. This is of particular importance because "real world data" is typically available without specific consent from the patient as a possible later use cannot be anticipated at the time of data collection.

Birth year and postal code are required to determine demographic and geographic distribution. This raises the question of whether the combination of age, location of residence, diagnosis, combined with an appointment date in a known clinic already represents an unacceptable deviation from anonymity. A method established to some extent to achieve a more complete blurring of all possible traces that could lead to an individual patient is changing of date information. Each data point exists together with a date stamp. From a data privacy perspective, it would be theoretically possibly to move one step further towards identification by using the date of a laboratory test. For this reason, some organizations obfuscate the meta data by intentionally changing the examination date or removing it entirely (chrononymization). For protocols with inclusion criteria that include a time frame (e.g. a specific diagnosis within the last 6 months, or a specific lab value in the last 12 weeks), this obfuscation significantly reduces the usefulness of the data. Studies on large volumes of records have shown that this method provides a false sense of security and that the availability of time information actually does not facilitate re-identification [12]. For this reason, this type of anonymization and chrononymization should be abandoned. It reduces at best the scientifically usable content of the information, without increasing data privacy protection.

The best solution to minimize data privacy concerns is the decentralized archiving of data (federated search). In this method, data remains permanently at the institution and is subject to the customary local protection mechanisms. External queries are processed in the institutional database or a local copy (within the firewall) and only the statistical result of a cohort (e.g. count, mean, standard deviation) is shared externally. Outside of the institution's own internal firewall, there is no additional database with sensitive information that would need protection or could potentially be placed on a server outside the individual legislation. In such a federated network, only statistical query results are consolidated from the individual member institutions, while the individual personalized records remain behind the firewall at the institution.

Pharm. Ind. 79, Nr. 2, 204–210 (2017)

■ Long-Term Model of Cooperation

Networking and cooperation work in the long term only if there is a driving motivation behind them. This can be a clear scientific objective, whereby the constant advance towards this objective must be recognizable. A commercial motivation is often the major driver. There must always be a good balance between benefits and costs/efforts. Otherwise, the enthusiasm of achieving such cooperation that is always present at the beginning will be replaced with complacency and the cooperation will subside.

In the case of health data, the data from the service providers (e.g. clinics) will be made available and one user group predominantly benefits. For example, why should a university hospital allow access to their data when others, such as the pharmaceutical industry, will benefit from it? This only works, if the benefits are equally distributed. The system must provide an equivalent advantage to the university hospital, e.g. use the platform for its own purposes, such as investigator-initiated studies or setup of its own networks, and therefore allowing data exchange not only with the pharmaceutical industry, but also with other institutions. Alternatively, reimbursement of costs to the clinic could be considered. However, the use of patient data for commercial, instead of scientific, purposes may shift the balance in data privacy discussions.

Conclusion

The clinical development of drugs is a highly data-oriented process. However, in clinical trials, the approach of statistical evidence is used only for providing proof of safety and efficacy. Other aspects and decision-making processes, for example the description of a target population for the desired indication, or the generation of protocols and selection of study sites, are still predominantly determined by (expert) opinions, prior subjective experience or semi-quantitative methods. They rarely are confirmed by real-world data from actual patients in routine medical care. The use of electronic health data can greatly contribute to developing more realistic protocols, in which inclusion and exclusion criteria are tested on real-world data at the design stage. A more targeted selection of sites, based on the availability of patients and defined by specific protocols, can help prevent an activation of sites that can provide no or only very few patients for the trial. Finally, the screening of patients is considerably easier if a protocol-specific query and re-identification of potential study patients in the site's database is possible.

The topic of big data obviously generates many questions, particularly about data protection in health

■ **Table 1**

Requirements Specification (score card) for comparing different providers.

Requirement	Priority	Provider 1	Provider 2	Provider 3
System can be used for:				
• Protocol design, cohort analysis				
• Feasibility testing of inclusion and exclusion criteria				
• Site selection				
• Patient pre-screening				
• Collaboration models across sites				
Completeness, i.e. contains:				
• Diagnoses				
• Medication				
• Procedures				
• Lab values				
• Findings				
• Genomics				
Representability:				
• Number of patients				
• Number of institutions providing data				
Up-to-Date:				
• Frequency of updates				
Site information:				
• Patient numbers over time				
User-friendliness:				
• Graphical user interface				
• Usability of Boolean logic variables				
• Linking with time frames/temporal events				
Data privacy:				
• Anonymization, chrononymization				
• Federated network vs. database				
• Interoperability				

care, but also about the magnitude of possibilities. This article targets a user audience, and therefore, it cannot go into more detail regarding technical requirements. Such requirements are discussed extensively in literature [13]. To help retain a clear view of the market, a Requirements Specification is recommended (Tab. 1). Data protection, usability across the entire process, volume and completeness of data, freshness of data, representability of data, geography as well as user-friendliness are the most important elements for a risks/benefits analysis of such a system for users in the clinical development environment.

LITERATURE

[1] BG O'Sully. A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies. Trials. 2013; 14: 166.

[2] CD Kohl. Patientenübergreifende, multiple Verwendung von Patientendaten für die klinische Forschung unter Nutzung von Archetypen. Inauguraldissertation zur Erlangung des Doctor scientiarum humanarum an der Medizinischen Fakultät Heidelberg der Ruprecht-Karls-Universität. 2012.

[3] K Getz et al. The Impact of Protocol Amendments on Clinical Trial Performance and Cost. Therapeutic Innovation & Regulatory Science. July 2016; Vol. 50, 4: 436–441.

[4] B Harper. Effective Strategies for Patient Recruitment and Retention. UC Davis Clinical and Translational Science Center. December 5, 2014.

[5] JW London et al. Design-phase prediction of potential cancer clinical trial accrual success using a research data mart. J Am Med Inform Assoc. 2013; 20: e260–e266.

[6] M Stapff. Quality by Design aspects applied to Clinical Study Protocol Development. Chimica Oggi – Chemistry Today. November/December 2014; Vol. 32(6).

[7] C Parke et al. The Louisiana Clinical Data Research Network: Leveraging Regional and National Resources to Improve Clinical Research Efficiency. The Ochsner Journal. 2014; 14: 718–723.

[8] B Trinczek et al. Design and multicentric Implementation of a generic Software Architecture for Patient Recruitment Systems re-using existing HIS tools and Routine Patient Data. Appl Clin Inform. 2014; 5: 264–283.

[9] WQ Wei et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. Am Med Inform Assoc. 2016; 23: e20–e27.

[10] RH Perlis et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med. January 2012; 42(1).

[11] M Stapff. Use of Electronic Health Records for Development and Feasibility Testing of Clinical Trial Protocols. DIA 28th Euro Meeting, April 2016, Hamburg, Germany.

[12] JJ Cimino. The False Security of Blind Dates. Appl Clin Inf. 2012; 3: 392–403.

[13] J Chahboune. IT-Unterstützung vs. Datenschutz im Bereich klinischer Studien. Pharm Ind. 2015; Vol. 77(2): 173–180.

Correspondence:
Dr. med. Manfred Stapff
SVP and Chief Medical Officer
TriNetX Inc.
125 Cambridgepark Drive, Suite 203
Cambridge, MA 02140 (USA)
e-mail: Manfred.Stapff@trinetx.com

Editor-in-chief: Claudius Arndt. Managing Secretary's office: Gudrun Below. Publisher: ECV · Editio Cantor Verlag für Medizin und Naturwissenschaften GmbH, Baendelstockweg 20, 88326 Aulendorf (Germany). Phone: +49 (0) 75 25 94 00, Fax: +49 (0) 75 25 94 01 80. e-mail: redaktion@ecv.de. http://www.ecv.de.
Typesetting: Reemers Publishing Services GmbH (Germany) / Printed by Holzmann Druck GmbH & Co. KG.