1

Clinical Data Collection and Patient Phenotyping

Katerina Markoska[1] *and Goce Spasovski*[2]

[1] *Medical Faculty, University "Ss. Cyril and Methodius" of Skopje, Skopje, Republic of Macedonia*
[2] *Department of Nephrology, University "Ss. Cyril and Methodius," Medical Faculty, Skopje, Republic of Macedonia*

1.1 Clinical Data Collection

1.1.1 Data Collection for Clinical Research

The goal of clinical studies is the evaluation of interventions on clinically relevant parameters [1]. Conducting a clinical study is a major undertaking accompanied with heavy and extensive responsibilities. Good primary research calls for constant dedication by practicing physicians and patients willing to participate for the sake of knowledge and better treatment of future patients [2].

The study design is the investigator's map from which data collection follows and which enables the investigators to thoughtfully produce the necessary data forms. The formulation of a good research question, up front, informs the clinician or researcher about the most appropriate data elements to be collected [2]. Investigators often believe that collecting more data is better and that it is important to collect information on as many scientifically "interesting" factors as possible. Therefore, it is imperative to distinguish between those data elements that are essential and those that are academically "interesting" but may not be considered of interest to the key study hypothesis. This should greatly assist in narrowing down one's study questions and collecting data more efficiently [3].

1.1.2 Clinical Data Management

Clinical data management (CDM) is the process of collection, cleaning, and management of subject data in compliance with regulatory standards. The primary objective of CDM processes is to provide high-quality data by keeping the number of errors and missing data as low as possible and gather the maximum amount of data for analysis [4].

High-quality data should be absolutely accurate, have minimal or no missing points, and should be suitable for statistical analysis. The data should meet the applicable regulatory requirements specified for data quality and comply with the protocol requirements. In case of a deviation or not meeting the protocol specifications, we may think of excluding the patient from the final database [5].

Current technological developments have accelerated the rate of data collection and positively impacted the CDM process and systems by improving their quality. From the regulatory perspective, the biggest challenge would be the standardization of data management processes across organizations and development of regulations to define the procedures that has to be followed. From the industry perspective, the challenge would be the planning and implementation of data management systems in a changing operational environment. CDM is evolving to become a standard-based clinical research entity, balancing between the expectations from and constraints in the existing systems, driven by technological developments and business demands [5].

The Society for Clinical Data Management (SCDM) publishes Good Clinical Data Management Practices (GCDMP) guidelines that highlight the minimum standards and best practices, providing assistance to clinical data managers in their implementation of high-quality CDM [5]. If data have to be submitted to regulatory authorities, it should be entered and processed in accordance with the Code of Federal Regulations (CFR), Title 21, Volume 1 of Part 11, Food and Drug Administration (FDA) regulations on electronic records and electronic signatures (ERES), cited as 21CFR11.10 [6].

Many clinical data management systems (CDMS) are available for data management. Most of the CDM systems available meet these criteria, and pharmaceutical companies as well as contract research organizations

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

ensure this compliance. In multicentric trials, a CDMS has become essential to handle the huge amount of data. These CDM tools ensure the audit trail and help in the management of discrepancies [5].

One should leave sufficient time for planning and development of system and study database for the follow-up and tracking of patients throughout the study. The following issues should be defined in advance: determination of the mechanism and processes for data collection if a patient misses a scheduled appointment, implementation of quality checks, preparation and collection of patient informed consent, and institutional review board (IRB) approval. Inclusion and exclusion criteria should be defined as well as data collection elements [2].

1.1.3 Creating Data Forms

The limited focus of disease-specific consortia makes comprehensive coverage of individual areas more likely. Researchers should benefit from a clear understanding of the extensive overlap of various clinical terminologies, as well as advice regarding which standards are appropriate for a particular research context. They should also be able to address relationships between clinical research data collection standards and electronic health records (EHR) specifications, as well as the broad issue of secondary use of clinical data for research. Additional tasks could include the review of standards and their scope and relating them to needs of clinical research [7].

Item repositories can reduce the burden on new investigators to create their own items, because existing validated items or sets of items can be reused [8]. Pilot testing of data forms completed by patients allows investigators to react to suggestions from patients as well as from staff and personnel and provides more realistic estimates of data collection times [2].

1.1.3.1 Different Data Forms According to the Type of Study

Data form development is a collaborative effort among the investigators and often takes months of planning and preparation. It should be undertaken by investigators and/or stakeholders experienced in form construction and familiar with the methods of data collection, data processing, and content necessary for the study [2]. It is facilitated by review of the literature for instruments used in similar studies, also including the Clinical Data Acquisition Standards Harmonization (CDASH) recommendations, which give useful general guidance on constructing yes/no questions, scale direction, date/time formats, scope of CRF data collection, pre-populated data, and collection of calculated or derived data. Certain items (especially questionnaire-based ones) have a discrete set of permissible values (also called "responses" or "answers"), for example, the use of cigarettes (never/former/current), amount (less than 3 per day/3–10 per day/more than 10 per day), and fasting (no/yes) [7].

Study details like objectives, intervals, visits, investigators, sites, and patients should be defined in the database, and CRF layouts have to be designed for data entry [5]. In order to simplify the data collection, some answers can be coded. For example, 1 = yes and 2 = no, but these codes should be consistent throughout the CRF booklet (Table 1.1) [9].

The forms should be well designed in order to avoid variation in the responses and the site personnel can understand the format (Table 1.2) [9].

Much of the information collected in observational epidemiologic studies is collected in the form of patient/participant self-reports on standardized questionnaires that are self-administered or administered in person by an interviewer, by phone, or via mail or the Internet. The factors on which information is routinely collected in these studies include sociodemographic characteristics, lifestyle practices, medical history, and use of prescribed or over-the-counter medications [3].

Surveys are tools of great value for epidemiological research and clinical practice. They can be used as a study design, at same time serving as definitive data collection tool. On the other hand, clinical registries

Table 1.1 Coding on the case report form module.

Demography	
Date of birth (DD/MM/YYYY)	☐☐/☐☐/☐☐☐☐
Gender	Male ☐ 1 Female ☐ 2
Height (cm)	☐☐☐.☐
Weight (kg)	☐☐☐.☐☐
Smoker	Yes ☐ 1 No ☐ 2
Family history	Yes ☐ 1 No ☐ 2

Table 1.2 Well-designed and poorly designed data fields.

Poorly designed	Well designed
Date of visit:_____	Date of visit: ☐☐/☐☐/☐☐☐☐ (DD/MM/YYYY)
Blood pressure:____/____	Blood pressure: ☐☐☐/☐☐☐ (mmHg)
Pulse: _____	Pulse: ☐☐☐ (beats/min)
Temperature:_____	Temperature: ☐☐.☐ (°C)
Respiration: _____	Respiration: ☐☐ (/min)

can be also used to obtain specific data within a more comprehensive design [10].

Ideally, patient-reported outcomes are measured using standardized, validated instruments that promote the collection of high-quality data and allow meaningful comparisons across observational studies or randomized trials. National Institutes of Health Toolbox (www.nihtoolbox.org) and Patient-Reported Outcomes Measurement Information System (www.nihpromis.org) have highlighted the importance of harmonization of patient-reported outcomes data collection instruments [3].

Clinical Document Architecture templates—archetypes—are agreed-upon specifications that support rigorous computable definitions of clinical concepts like type of measure, measurement conditions, and measurement units [11].

OpenEHR, by contrast, allows the semantic model's structure to vary with the parameter being described. Clinical researchers have been specifying parameter measurement with precision long before "archetypes" were conceived [7].

CDASH addresses data collection standards through standardized CRFs. Initial CDASH standards focused on cross-specialty areas such as clinical trial safety.

The Clinical Data Interchange Standards Consortium (CDISC) (http://www.cdisc.org) is an international standards organization that aims to develop and support global, platform-independent data standards. The consortium has proposed standards valuable for general areas such as drug safety, focusing primarily on regulated studies, and does not address broader issues of clinical research. They also have CDISC Operational Data Model (ODM) for exchanging and archiving clinical study data [12]. Another clinical information model is the Health Level 7 (HL7) Reference Information Model (RIM), which use terminologies differently than the research-oriented CDISC (ODM) format. HL7 depends on mapping data elements to concepts in standard terminologies, while ODM does not support mapping of data elements themselves (e.g., serum total cholesterol, systolic BP) to terminologies and only cares that a terminology may act as a source for a data element's contents [13].

1.1.4 Case Report Form (CRF)

1.1.4.1 CRF Standards Characterization

Data collection for clinical research involves gathering variables relevant to research hypotheses. These variables ("patient parameters," "data items," "data elements," or "questions") are incorporated into data collection forms ("case report forms" (CRFs)) for study implementation [4].

CRF may exist in the form of a paper or an electronic version. The traditional method is to employ paper case report forms (pCRFs) to collect the data responses, which are translated to the database. These pCRFs are filled up by the investigator according to the completion guidelines. In the electronic case report form (eCRF)-based CDM, the investigator or a designee will be entering the data directly at the site. In the case of eCRF, the probability of erroneous data entry is lower, and the resolution of discrepancies faster [5].

A CRF is designed by the CDM team, as this is the first step in translating the protocol-specific activities into data being generated. The units in which measurements have to be made should also be mentioned next to the data field [5].

Because of the protocol-centric nature of clinical research, opportunities for shared standards at levels higher than individual items are relatively limited. Nevertheless, disease-specific CRF standardization efforts have helped identify standard pools of data items within focused research and professional communities and consequently helped achieve research efficiencies within their application areas. Of more immediate and widespread relevance are standardization efforts toward the development of section and workflow for CRF, as well as data collection and validation. The structure and content of individual CRFs/sections can be left reasonably flexible to allow adaptation to individual protocol requirements [7].

Little consensus exists on the choice and content of CRF standardization candidates. Few CRFs can be reused unchanged across all protocols. Within a specific disease domain, standard CRFs seem feasible and useful. But the segregation of data items relevant to a research protocol into individual CRFs is often based on considerations other than logical grouping and may vary with the study design. One concern about "standard" CRF use is that users should not be pressured to collect parameters defined within the CRF that are not directly related to a given protocol's research objectives. Dynamic CRF rendering offers one way out of this dilemma: protocol-specific CRF customization allows individual investigators to specify, at design time, the subset of parameters that they consider relevant. Also, web application software can read the customization metadata and render only applicable items [7].

Generally, a programmer/designer performs the CRF annotation, creates the study database, and programs the edit checks for data validation. He/she is also responsible for designing of data entry screens in the database and validating the edit checks with dummy data [5]. Databases are the clinical software applications, which are built to facilitate the CDM tasks to carry out multiple studies [14].

CRFs can be used in groups of semantically closely related parameters, which can be considered as a series

of observations. A section encompasses one or more groups. The division of CRFs into sections is often arbitrary. In paper-based data entry, CRFs consisting of a single, oversize section are sometimes used. Real-time electronic data capture (EDC) subdivision into smaller sections is generally preferred. Section headings and explanations that serve to describe the section's purpose are usually left to individual investigators [7].

1.1.4.2 Electronic and Paper CRFs

CRFs support either primary (real-time) data collection or secondarily recorded data originating elsewhere (e.g., the electronic or paper health records). Historically, CRFs were paper based. The existence of secondary EDC also influences manual workflow processes related to verification of paper-based primary data (e.g., checks for completeness, legibility, and valid codes) [7].

Although the use of validated, standardized instruments is preferred, those data collection tools are not always available. If standardized instruments do not exist for measuring a specific construct, investigators will often create "homegrown" scales, which require pilot test before using them in a formal research study. These pilot efforts ideally would involve validation of the instrument against a gold standard (e.g., clinical diagnosis) or important study outcome [3].

Collection of individual patient data on CRFs in clinical research has traditionally been done by investigators in their offices summarizing medical charts on paper forms (pCRFs), a tedious method that could result in data entry errors and wrong conclusions [15, 16].

eCRFs have improved data quality and completeness, reducing losses and transport logistics, especially for multicenter trials [17]. The choice between pCRF and eCRF is a significant step in the design of clinical studies and should be discussed with the involved stakeholders [18].

EHR and research data collection differ in that the latter records a subset of patient parameters and variables defined with the research protocol. Data are recorded in maximally structured form, avoiding narrative text, except if there is a need to record unanticipated information [7].

Le Jeannic et al. have compared the application of eCRF and pCRF and their results showed that eCRF studies were mostly used in large multicenter, national, and phase 3 clinical trials while pCRF studies were used for trials with few patients and centers. The majority of pCRFs were used in drug trials, and eCRFs were more often used in trials with a significantly higher number of patients and fewer data. The number of patients was the only explanatory variable for CRF choice. They found no difference in the average duration of recruitment. Use of eCRF and the smaller number of centers were associated with shorter study durations. The total average cost of a trial was higher with eCRFs than with pCRFs, but the mean cost per patient was lower with eCRFs. Overall, stakeholders were as satisfied with eCRFs as with pCRFs. When asked for their preference of one over the other, a majority of stakeholders chose eCRF. Preference for pCRFs is reported in monocentric trials and for eCRFs in multicentric trials. Additional advantages of eCRFs are the prevention of data entry errors by automatic checks, easier storage, and the ability of researches to oversee data collection from their offices [18].

1.1.5 Methods and Forms for Clinical Data Collection and/or Extraction from Patient's Records

In particular, a patient summary has been seen as the most appropriate way to establish eHealth interoperability. A patient summary includes patient history, allergies, active problems, test results, and medications. However, further information can be included, depending on the intended purpose of the summary and the anticipated context of use [19].

Because of the ubiquity and abundance of high-quality data embedded within medical records, they are a commonly used source of information in clinical research studies. Medical records can be important sources of information that can reliably document participants' medical history, clinical, laboratory, or physiologic profile at varying time points in a cost-efficient manner. On the other hand, the data contained in medical records can be difficult to use and, in some cases, conflicting or of questionable accuracy because of the nonstandardized manner in which this information is collected, recorded, and extracted by various healthcare professionals and members of research teams. The increasing use of electronic medical records (EMR) and their combination with administrative data have eased data extraction efforts. Moreover, the increasing use of standardized data entry sets reduced data heterogeneity [3].

1.1.5.1 Electronic Health Records (EHRs)

EHRs are basically seen as a centralized compilation of information on the patient's health [20]. The data included in paper-based patient records has provided the golden standard against which the reliability of EHRs has been assessed. The success of EHRs depends on the quality of the information available to healthcare professionals in making decisions about patient care and in the communication between healthcare professionals during patient care [19]. It has been shown that data from EHRs are reliable when compared with manual records [21, 22].

One challenge is to standardize health information systems, which also means standardization of the content and structure of EHRs [23]. EHRs have so far consisted

of unstructured narrative text but also contain structured coded data [19]. EHR contains retrospective, concurrent, and prospective information, and its primary purpose is to support continuing, efficient, and quality-integrated healthcare [24]. An EHR is used primarily for purposes of setting objectives and planning patient care, documenting the delivery of care, and assessing the outcomes of care. It includes information regarding patient needs during episodes of care provided by different healthcare professionals [25, 26]. The EHR is used by different healthcare professionals and also by administrative staff. Among the various healthcare professionals who use different components of the EHR are physicians, nurses, radiologists, pharmacists, laboratory technicians, and radiographers [19]. EHRs are used by many different healthcare professionals, and the needs and requirements of all these professionals must be taken into account in the development of the information systems. Nursing documentation, or documentation by other healthcare professionals such as physiotherapists, is an important part of the EHR and must not be excluded from medical documentation. Patients can also do parts of the documentation themselves. Patient self-documentation also reduces the workload of healthcare professionals, but it is obviously important that self-documented data components are validated by professionals [19].

Previously EHRs were classified as time oriented, problem oriented, and source oriented. Nowadays EHRs combine all three elements. In the time-oriented EMR, the data are presented in chronological order. In the problem-oriented medical record, notes are taken for each problem assigned to the patient, and each problem is described according to the subjective information, objective information, assessments, and plan. In the source-oriented record, the content of the record is arranged according to the method by which the information was obtained, for example, notes of visits, X-ray reports, and blood tests [19].

Electronic clinical records, such as conventional clinical histories, can display major shortcomings in terms of quality of information, lack of data, incomplete information, and use of multiple free terms. Before electronic clinical records can replace registries or surveys, a common terminology and set of standards must be established to encode and classify the information, and a change must be brought about in the attitude of health professionals tasked with data collection [10].

The possibility of using electronic clinical histories as a data source may depend upon the degree to which this is used within the health organization and/or system (sole data collection source, data also recorded in paper format, etc.), the completeness and coding of recorded data, and also the software available for data collection and transfer [27].

Introducing an online medical record system could play an important role in improving data collection and data quality [28].

1.1.6 Data Collection Workflow

1.1.6.1 Defining Baseline and Follow-Up Data

Before data collection begins, investigators must agree on the details of the data collection items and the process by which data collection will occur. Investigators must define the schedule according to which patients will participate in the study and outline the specific data elements to be collected each time the patient is examined. If the researcher understands office flow and can organize the follow-up process, then his or her office can map data collection in a simple and efficient manner [2].

In most cases, it is best to collect all the required initial data for a subject during a single visit at the clinic. Several steps and design features are recommended to optimize follow-up rates [2].

In order to minimize the respondent burden, follow-up questionnaires and tests should be kept to a minimum. Contact information should be collected at baseline and updated at every visit for data collection, whereas subjects with no telephone or who plan to move in the near future should be excluded. Clinicians need to plan multiple efforts at phone contact, both during and after working hours, and provide reminders for appointments. Follow-up forms should include information about treatment compliance and the exposure of patients to various operative and nonoperative treatments. Follow-up forms must also include data regarding side effects and complications of treatment (e.g., monitored events) and whether they are related to the study treatment(s). In addition to baseline and follow-up patient data, information regarding treatment must be collected [2].

1.1.6.2 Medical Coding

Pre- or coexisting illnesses are coded using the available medical dictionaries. Medical Dictionary for Regulatory Activities (MedDRA) is used for the coding of adverse events as well as other illnesses. The World Health Organization Drug Dictionary Enhanced (WHODDE) is used for coding the medications. Medical coding helps in classifying reported medical terms on the CRF to standard dictionary terms in order to achieve data consistency and avoid unnecessary duplication. The right coding and classification of adverse events and medication is crucial as an incorrect coding may lead to masking of safety issues or highlight the wrong safety concerns related to the drug [5].

It also is important to note that factors (e.g., medication use) must be defined only by clinicians, and not by study staff or study participants, in order to ensure that variables will be accurately coded [3].

1.1.6.3 Errors in Data Collection and Missing Data

If the data are inaccurate or incomplete, they will have no worth for decision-making, research, statistical, or health policy purposes [15].

One common cause of errors is not dedicating enough time in the development of data forms (i.e., in identifying the data elements and in the construction and testing of forms). An important problem is the desire of clinicians not only to create forms to meet the research goals of the study but also to provide data for routine patient care. Certain measurements needed for routine patient care are not justifiable for research forms, and vice versa. The researcher should be clear about the data necessary for assessing the primary and secondary outcomes of the study. Data items that cannot be justified should be deleted [2].

Double data entry is performed wherein the data is entered by two operators separately. The second pass entry helps in data verification by identifying the transcription errors and discrepancies caused by illegible data. Double data entry helps in getting a cleaner database compared with a single data entry. Double data entry ensures better consistency with pCRF as denoted by a lesser error rate [29, 30].

It is difficult to gather the necessary data elements at the appropriate times while avoiding missing data. It is even more difficult to collect primary data according to a very strict protocol, wherein chance, bias, and confounding factors can be addressed. No matter how sophisticated the data elements and data collection systems, human factors make or break any good research effort [2].

With increasing duration of a study, the number of participating patients usually declines, so the problem of missing data is magnified. Thus, data interpretation for long-term studies is challenging [31].

A frequently employed approach for data analysis of clinical (long-term) studies is the interpretation of missing data as therapeutic success (missing equals success (MES)) or as therapeutic failure (missing equals failure (MEF)/nonresponder imputation). A third option the exclusion of missing data (missing equals excluded (MEX)/as-treated) stands between these two extremes. Another frequently employed method for long-term studies that is criticized by statisticians is equating the last observed value with the result at the end of the study (last observation carried forward (LOCF)). The selection of the analysis method has a great impact on the results and interpretation of a study. It is recommended to combine several data analysis approaches in order to correctly interpret long-term studies and reach valid conclusions. A comparison of the characteristics of test subjects with complete as opposed to those with incomplete datasets might be helpful, in order to get indications on the possible reasons for dropping out of the study or for missing data [1].

1.1.6.4 Data Linkage, Storage, and Validation

The data management group (those responsible for data retrieval and processing) needs to link records by using a unique identifier for each patient. For example, they might use the patient hospital identification (for purposes of confidentiality, patient identification can be deleted later) plus check digits that identify the patient, the center from which the patient comes, and the type of visit [2].

The data bank must be backed up regularly. Data collection is too difficult and expensive to repeat, and patients' time is too valuable to have data lost or destroyed [2].

Data validation is the process of testing the validity of data in accordance with the protocol specifications. Discrepancy is defined as a data point that fails to pass a validation check, and it may be due to inconsistent data, missing data, range checks, and deviations from the protocol. Ongoing quality control of data processing is undertaken at regular intervals during the course of CDM. Data clarification forms (DCFs) containing queries pertaining to the discrepancies identified can be also generated [5].

In order to prevent errors from being entered, data validation rules should be implemented into the eCRF's prior to commencement of the NPC clinical trial. These data validation rules assess whether certain prespecified conditions are valid and can therefore pinpoint omissions or erroneous data [28].

The clinical trial data management system (CTDMS) prevents us from missing data or ending up with poor quality data at the end of the study, which often at that point cannot be resolved anymore [28].

Clearly, the CTDMS encourages local data managers to verify the entered data and, if necessary, ask the doctor whether the information is correct [28].

Discrepancy management helps in cleaning the data and gathers enough evidence for the deviations observed. Discrepancy management is the most critical activity in the CDM process. Being the vital activity in cleaning up the data, utmost attention must be observed while handling the discrepancies [5].

After a proper quality check and assurance, the final data validation is run. Database is locked and clean data is extracted for statistical analysis. Data extraction is done from the final database after locking. This is followed by its archival [5].

1.2 Patient Phenotyping

In clinical care settings, a wealth of longitudinal data is available through International Statistical Classification of Diseases v9 (ICD-9) codes, laboratory results,

test reports, and notes written by the physicians during multiple patient visits over several years. Essential point in improving the formation of research cohorts has been the creation of EMR-linked biobanks and enrolment of individuals from routine clinical care settings. With patient's consent and by making their data anonymous, EMRs can make a large amount of information available for research purposes, allowing studying the evolution and progression of the disease [32–34]. In this regard, using large number of patient's data from EMRs can markedly reduce the time and effort needed to identify specific phenotypes and/or markers associated with disease development, progression, and response to treatment [34–36].

1.2.1 Approaches in Defining Patient Phenotype

The Electronic Medical Records and Genomics (eMERGE) Network is a National Human Genome Research Institute (NHGRI)-funded consortium tasked with developing methods and best practices for the utilization of the EMR as a tool for genomic research. The network is developing phenotyping algorithms that are processing EMR data in order to identify cases and controls with a high degree of accuracy and confidence [37]. However, identification of particular phenotypes, especially chronic complex diseases, is challenging because of the complexity of data itself and the way in which it is recorded in EMR [38].

There has been significant debate about the optimal way to identify phenotypes in the EMR. Automated approaches using electronic phenotyping and statistical analyses are popular as compared with simpler rule-based systems. Such phenotyping algorithms are used in various applications including discovering novel genetic associations of complex diseases, tracking their natural history, isolating patients for clinical trials, and ensuring quality control in large institutions by ensuring that standard-of-care guidelines are met in these patients [38, 39].

Phenotyping algorithms are dedicated to mining biobank resources, which is essential for trial designs, and need to enable automatic identification of patients that match the research criteria. They contain keywords designed to facilitate natural language processing (NLP) and access the primarily semi-structured data fields in EHRs—procedure codes, ICD-9 codes, laboratory results, and medication data [40].

NLP content is required for enormous unstructured narrative clinical documentation that is considered to be the best resource. This is the most difficult part in phenotype algorithm construction, and although there are many NLP tools for medical domains, human involvement is still required [40].

The V-Model is a temporal model that enables visualization of textual information in a timeline, which helps in monitoring and understanding a patient's history. The model separately structures causal problems from related actions, representing apparent problems–actions (P-A) relations, and enables to extrapolate that problems occurred before actions. It enables a user to trace patient history considering semantic, temporal, and causality information in a short time. Consequently the V-Model should play a crucial role in phenotype definition and algorithm development [41].

For several relatively common conditions, such as heart failure and stroke, independently and extensively validated algorithms have been developed to ascertain the presence of these important chronic diseases [42]. eMERGE Network developed 14 robust algorithms that were extensively tested over multiple iterations (Table 1.3). The core elements of the algorithms are the administrative data (ICD-9 and CPT codes), laboratory data, and medication data (RxNorm codes), with NLP rules as an additional layer to disambiguate and refine the core data elements. Most of the developed algorithms rely on ICD-9 disease codes and use CPT procedure codes. Only two algorithms used UMLS codes (due to site-specific processing needs) [40].

1.2.2 Phenotyping CKD Patients

Clinical decision making is challenging due to variability in the rates of progression and lack of widely accepted guidelines to identify patients most at risk of progression to ESRD [43, 44].

Currently the only way of identifying CKD cases/controls is by manually reviewing laboratory values, which is cumbersome, or through ICD-9 codes. To accomplish these goals, researchers need robust phenotyping algorithms to effectively leverage disparate data sources in the EMR [38].

Nadkarni et al. developed and validated an automated algorithm for identifying diabetic/hypertensive CKD cases and controls. Their algorithm over-performed the traditional identification using ICD-9 diagnostic codes, which enabled identification of 40.1% of cases and 75.0% of controls. Their algorithm correctly identified 93.4% of cases and 95.8% of controls, indicating that it could be used for both research and clinical purposes, where rapid and accurate identification of a specific target cohort is needed [38].

Table 1.3 Distribution of codes across the 14 eMERGE phenotyping algorithms.

	Name	ICD-9[a]	CPT[b]	UMLS[c]	RxNorm[d]	Total/WC[e]	Percentage[f]
1	Alzheimer's	29	0	0	355	384/1 317	29
2	Dementia	30	0	0	20	50/634	7
3	Diabetic retinopathy	12	19	0	0	31/324	10
4	Height	156	0	0	11	167/2 101	8
5	Hypothyroidism	43	76	0	0	119/1 351	9
6	Serum lipid level	11	0	0	0	11/1 091	1
7	Low HDL[g] cholesterol level	41	10	0	0	51/2 579	2
8	Peripheral arterial disease	90	112	0	0	202/1 353	15
9	QRS duration	50	157	595	0	802/26 695	3
10	Red blood cell indices	146	141	0	0	287/2 857	10
11	Resistant hypertension	35	0	0	0	35/895	4
12	Type 2 diabetes	25	0	0	0	25/954	3
13	White blood cell indices	18	131	0	0	149/2 458	6
14	Cataract	152	20	35	0	207/3 152	6

[a] Number of ICD-9 (International Statistical Classification of Diseases, v9) codes present in the algorithm document.
[b] Number of CPT (Current Procedure Terminology) codes in document.
[c] Number of UMLS (Unified Medical Language System) codes in the document.
[d] Number of RxNorm (clinical drug) codes in the document.
[e] Total number of codes divided by the number of word tokens.
[f] Percentage of the document's word tokens that are codes.
[g] High-density lipoprotein.

1.3 Concluding Remarks

Recommendations for standardization of CRF and transition from paper to electronic health records have significantly improved and accelerated the process of clinical data collection. Technological developments in the field of CDM have enabled fast, accurate, and simplified extraction of information from enormous clinical documentations for best quality of data generated. NLP tools that are used for extracting unstructured narrative clinical data and EHR-oriented phenotyping algorithms should enable automatic selection of cases for clinical trials and other research purposes.

References

1 Boehncke, W.H., Clinical long-term studies: data collection and assessment. *J Dtsch Dermatol Ges*, 2011. 9(6): p. 479–484.
2 Weinstein, J.N. and R.A. Deyo, Clinical research: issues in data collection. *Spine*, 1976. 25(24): p. 3104–3109.
3 Saczynski, J.S., D.D. McManus, and R.J. Goldberg, Commonly used data-collection approaches in clinical research. *Am J Med*, 2013. 126(11): p. 946–950.
4 Gerritsen, M.G., et al., Data management in multi-center clinical trials and the role of a nation-wide computer network. A 5 year evaluation. *Proc Annu Symp Comput Appl Med Care*, 1993: p. 659–662.
5 Krishnankutty, B., et al., Data management in clinical research: an overview. *Indian J Pharmacol*, 2012. 44(2): p. 168–172.
6 Code of Federal Regulations Title 21. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=11.10 (accessed August 24, 2017).
7 Richesson, R.L. and P. Nadkarni, Data standards for clinical research data collection forms: current status and challenges. *J Am Med Inform Assoc*, 2011. 18(3): p. 341–346.
8 Brandt, C.A., et al., Approaches and informatics tools to assist in the integration of similar clinical research questionnaires. *Methods Inf Med*, 2004. 43(2): p. 156–162.
9 Bellary, S., B. Krishnankutty, and M.S. Latha, Basics of case report form designing in clinical research. *Perspect Clin Res*, 2014. 5(4): p. 159–166.

10 Varela-Lema, L., A. Ruano-Ravina, and T. Cerda Mota, Observation of health technologies after their introduction into clinical practice: a systematic review on data collection instruments. *J Eval Clin Pract*, 2012. 18(6): p. 1163–1169.

11 Browne E. Archetypes for HL7 CDA Documents, 2008. https://openehr.atlassian.net/wiki/spaces/stds/pages/5373955/openEHR+Archetypes+for+HL7+CDA+Documents?preview=/5373955/5537794/Archetypes_in_CDA_4.pdf (accessed October 12, 2017).

12 CDISC. Clinical Data Acquisition Standards Harmonization: Basic Data Collection Fields for Case Report Forms. Draft version 1.0, 2008.

13 Richesson, R.L. and J. Krischer, Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc*, 2007. 14(6): p. 687–696.

14 Fegan, G.W. and T.A. Lang, Could an open-source clinical trial data-management system be what we have all been looking for? *PLoS Med*, 2008. 5(3): p. 0050006.

15 Day, S., P. Fayers, and D. Harvey, Double data entry: what value, what price? *Control Clin Trials*, 1998. 19(1): p. 15–24.

16 Nahm, M.L., C.F. Pieper, and M.M. Cunningham, Quantifying data quality for clinical trials using electronic data capture. *PLoS One*, 2008. 3(8): p. 0003049.

17 Thwin, S.S., et al., Automated inter-rater reliability assessment and electronic data collection in a multi-center breast cancer study. *BMC Med Res Methodol*, 2007. 7: p. 23.

18 Le Jeannic, A., et al., Comparison of two data collection processes in clinical studies: electronic and paper case report forms. *BMC Med Res Methodol*, 2014. 14(7): p. 1471–2288.

19 Hayrinen, K., K. Saranto, and P. Nykanen, Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform*, 2008. 77(5): p. 291–304.

20 Friedman, D.J., Assessing the potential of national strategies for electronic health records for population health monitoring and research. *Vital Health Stat*, 2006. 2(143): p. 1–83.

21 Stratmann, W.C., A.S. Goldberg, and L.D. Haugh, The utility for audit of manual and computerized problem-oriented medical record systems. *Health Serv Res*, 1982. 17(1): p. 5–26.

22 Jamison, R.N., et al., Electronic diaries for monitoring chronic pain: 1-year validation study. *Pain*, 2001. 91(3): p. 277–285.

23 European Commission. e-Health—Making Healthcare Better for European Citizens: An Action Plan for a European e-Health Area, 2004.

24 ISO/DTR 20514, Health Informatics—Electronic Health Record—Definition, Scope, and Context, 2004. https://www.iso.org/standard/39525.html (accessed October 12, 2017).

25 van Ginneken, A.M., The computerized patient record: balancing effort and benefit. *Int J Med Inform*, 2002. 65(2): p. 97–119.

26 Grimson, J., Delivering the electronic healthcare record for the 21st century. *Int J Med Inform*, 2001. 64(2–3): p. 111–127.

27 Vlug, A.E., et al., Postmarketing surveillance based on electronic patient records: the IPCI project. *Methods Inf Med*, 1999. 38(4–5): p. 339–344.

28 Wildeman, M.A., et al., Can an online clinical data management service help in improving data collection and data quality in a developing country setting? *Trials*, 2011. 12(190): p. 1745–6215.

29 Cummings, J. and J. Masten, Customized dual data entry for computerized data analysis. *Qual Assur*, 1994. 3(3): p. 300–303.

30 Reynolds-Haertle, R.A. and R. McBride, Single vs. double data entry in CAST. *Control Clin Trials*, 1992. 13(6): p. 487–494.

31 Committee for Proprietary Medicinal Products (CPMP). Points to consider on missing data: the European Agency for the Evaluation of Medicinal Products (EMEA), 2001.

32 Blumenthal, D. and M. Tavenner, The "meaningful use" regulation for electronic health records. *N Engl J Med*, 2010. 363(6): p. 501–504.

33 McCarty, C.A., et al., Informed consent and subject motivation to participate in a large, population-based genomics study: the Marshfield Clinic Personalized Medicine Research Project. *Community Genet*, 2007. 10(1): p. 2–9.

34 Denny, J.C., et al., PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, 2010. 26(9): p. 1205–1210.

35 Carroll, R.J., A.E. Eyler, and J.C. Denny, Naive Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA Annu Symp Proc*, 2011. 2011: p. 189–196.

36 Birman-Deych, E., et al., Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care*, 2005. 43(5): p. 480–485.

37 Gottesman, O., et al., The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med*, 2013. 15(10): p. 761–771.

38 Nadkarni, G.N., et al., Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu Symp Proc*, 2014. 14: p. 907–916.

39 Shivade, C., et al., A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*, 2014. 21(2): p. 221–230.

40 Conway, M., et al., Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. *AMIA Annu Symp Proc*, 2011. 2011: p. 274–283.
41 Park, H. and J. Choi, V-Model: a new perspective for EHR-based phenotyping. *BMC Med Inform Decis Mak*, 2014. 14(90): p. 1472–6947.
42 Carnahan, R.M., Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. *Pharmacoepidemiol Drug Saf*, 2012. 21 (Suppl 1): p. 90–99.
43 Keane, W.F., et al., Risk scores for predicting outcomes in patients with type 2 diabetes and nephropathy: the RENAAL study. *Clin J Am Soc Nephrol*, 2006. 1(4): p. 761–767.
44 Keith, D.S., et al., Longitudinal follow-up and outcomes among a population with chronic kidney disease in a large managed care organization. *Arch Intern Med*, 2004. 164(6): p. 659–663.