# Chapter 11

# Electronic Health Records as Source of Research Data

**Wenjuan Wang, Davide Ferrari, Gabriel Haddon-Hill, and Vasa Curcin**

## Abstract

Electronic health records (EHRs) are the collection of all digitalized information regarding individual's health. EHRs are not only the base for storing clinical information for archival purposes, but they are also the bedrock on which clinical research and data science thrive. In this chapter, we describe the main aspects of good quality EHR systems, and some of the standard practices in their implementation, to then conclude with details and reflections on their governance and private management.

**Key words** Electronic health records, Data science, Machine learning, Data quality, Coding schemes, SNOMED-CT, ICD, UMLS, Data governance, GDPR

## 1 Introduction

Vast quantities of data are routinely recorded as part of the care process. While its primary aim is managing individual's patient care, there are significant opportunities to use these data to address research questions of interest. In the United Kingdom, there has been almost 25 years of research using routine primary care data, anonymized at source, through the General Practice Research Database (now CPRD, Clinical Practice Research Datalink [1]), and other data sources, also pooling data from multiple practices and tied to specific electronic health record (EHR) systems (QResearch [2], ResearchOne [3]). As better described in Subheading 4, we define anonymized data as one for which all elements that can link back to its owner are irrecoverably deleted; alternately there are pseudo-anonymization options that allow the reidentification of the owner through a procedure mediated by those responsible for that data security and privacy protection. Health Data Research UK has created a nationwide registry of EHR-derived datasets available for research [4]. A similar development has taken place in the Netherlands, where, in the early 1990s, the Netherlands Institute for Health Services Research (NIVEL) developed its Netherlands Information Network of General Practice [5], now named NIVEL

Olivier Colliot (ed.), *Machine Learning for Brain Disorders*, Neuromethods, vol. 197, https://doi.org/10.1007/978-1-0716-3195-9_11,

Primary Care Database (NIVEL-PCD) [6, 7]. Belgium also has its Intego Network [7, 8]. France has the *Système National des Données de Santé* [9, 10] and the data warehouse of *Assistance Publique-Hôpitaux de Paris (AP-HP)* [11]. Sweden has numerous and extensive nationwide registries [12]. These databases provide valuable information about the use of health services and developments in population health. In the United States, there has not been a tradition of using routine anonymized data, largely because the Health Insurance Portability and Accountability Act (HIPAA) regulations place restrictions on the linkage of health data from different sources without consent [13–15] and because small office practices have not been widely computerized. Instead, the focus has been mainly on secondary care (hospital) data, facilitated by the National Institutes of Health's (NIH) Clinical Translational Science Awards (CTSA) [16]. Use or reuse of administrative data for research purposes is becoming more restricted in Europe as well, partly as a consequence of the European General Data Protection Regulation (GDPR) that was established in 2016 [17, 18]. In addition, data owners increasingly want control over the use of their data, making it more difficult to construct large centralized databases.

## 2 Data Quality in EHR

An electronic health record (EHR) is a digital version of a patient's medical history which may include all of the key administrative clinical data relevant to that person's care, including demographics, vital signs, diagnoses, treatment plans, medications, past medical history, allergies, immunizations, radiology reports, and laboratory and test results. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. EHRs have been adopted with the aim of improving quality of patient care quality, in particular by ensuring that all pertinent medical information is being shared as needed for different care providers. Meantime, the rapidly growing number of EHRs has led to increasing interest and opportunities for various research purposes. To ensure the patients receive care as they need and to draw valid and reliable research findings, quality data are needed.

Data quality is defined as "the totality of features and characteristics of a data set that bear on its ability to satisfy the needs that result from the intended use of the data" [19]. Currently, there is no definitive agreement on the components of data quality in available research. Feder described in a study [20] frequently reported components of data quality including data accuracy (data must be correct and free of errors), completeness (data must be sufficient in breadth, depth, and scope for its desired use), consistency (data must be presented in a consistent format),

credibility (data must be regarded as true and credible), and timeliness (data should be recorded as quickly as possible and used within a reasonable time period) [20–24]. Other aspects of data quality might include accessibility which means that data must be available for use or easily retrievable, appropriate amount of data which means the quantity of data must be appropriate, ease of understanding which means data must be clear, interpretability which means data must be in appropriate language and units, etc.

Many concerns were raised on digital data quality within EHRs including incompleteness, duplication, inconsistent organization, fragmentation, and inadequate use of coded data within EHR workflows [25]. As the old programming maxim states: garbage in, garbage out. Poor data quality can impact the care patients receive which may in turn lead to long-term damage or even death. It will also impact public health decision-making whenever it is based on statistics drawn from inaccurate data. In the following section, we will investigate in more detail the challenges regarding data accuracy and data completeness.

### 2.1 Data Accuracy

Data accuracy can be conceptualized as how accurate or truthful the data captured through the EHR system is. In other words, it is the degree to which the value in the EHR is a true representation of the real-world value [20, 23, 24] (e.g., whether a medication list accurately reflects the number, dose, and specific drugs a patient is currently taking [21]). A pilot study evaluated information accuracy in a primary care setting in Australia and confirmed that errors and inaccuracies exist in EHR [26]. This pilot study showed that high levels of accuracy were found in the area of demographic information and moderately high levels of accuracy were reported for allergies and medications. A considerable percentage of non-recorded information was also present. The sources of data inaccuracy could be mistakes made by clinicians (e.g., clinicians improperly use the "cut and paste" function in electronic systems [27]), error, loss or destruction of data during a data transfer [27]. Ways to improve data accuracy at collection include avoiding EHR pitfalls (e.g., fine-tuning preference lists, being careful when copying data, modifying templates as needed, documenting what was done, etc.) and being proactive (e.g., conducting regular internal audits, training staff, maintaining a compliance folder, etc.).

Data accuracy can be assessed via different approaches [20]. One can compare a given variable within the dataset to other variables which is referred to as internal validity, e.g., using medication to confirm the status of the disease. Internal validation can also be done by looking for unrealistic values (a blood pressure that is too high or low [28]) which could be checked by identifying outliers. One can also use different data sources or datasets to cross-check the data accuracy which is referred to as external validity, e.g., a patient was registered in a stroke registry but recorded as not

having a stroke in the current dataset. Generally, it is hard to link multiple datasets due to data privacy policy. Simple statistical measures can help the researcher determine whether variable values follow logical restrictions and patterns in the data such as central tendency (e.g., mean, median, mode) and dispersion (range, standard deviation) for continuous variables and frequencies and proportions for categorical variables and goodness-of-fit tests (e.g., Pearson chi-square) [20]. Researchers found that validation helps check the quality of the data and identify types of errors that are present in the data [28].

### 2.2 Data Completeness

Data completeness is referred to as the degree and nature of the absence of certain data fields for certain variables or participants. Generally, these absent values are called missing data. Missing data is very common in all kinds of studies, which can limit the outcomes to be studied, the number of explanatory factors considered, and even the size of the population included [28] and thus reduce the statistical power of a study and produce biased estimates, leading to invalid conclusions [29]. Data may be missing due to a variety of reasons. Some data might not be collected due to the design of the study. For example, in some questionnaires, certain questions are only for females to answer which leads to a blank for males for that question. Some data may be missing simply because of the breakdown of certain machines at a certain time. Data can also be missing because the participant did not want to answer. Some data might be missing due to mistakes during data collection or data entry. Thus, knowing how and why the data are missing is important for subsequent handling and for analyzing the mechanism underlying missing data.

Depending on the underlying reason, missing data can be categorized into three types [30] (Fig. 1): missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR is defined as data to be missing not related to any other variables or the variable itself. Examples of MCAR are failures in recording observations due to random failures with experimental instruments. The reasons for its absence are normally external and not related to the observations themselves. For MCAR, it is typically safe to remove observations with missing values. The results will not be biased but the test might not be powerful as the number of cases is reduced. This assumption is unrealistic and hardly happens in practice. For missing data that are MAR, missingness is not random and can be related to the observed data but not to the value of this given variable [31]. For example, a male participant may be less likely to complete a survey about depression severity than a female participant [32]. The data is missing because of gender rather than because of the depression severity itself. In this case, the results will be biased if we remove patients with missing values as most completed observations are

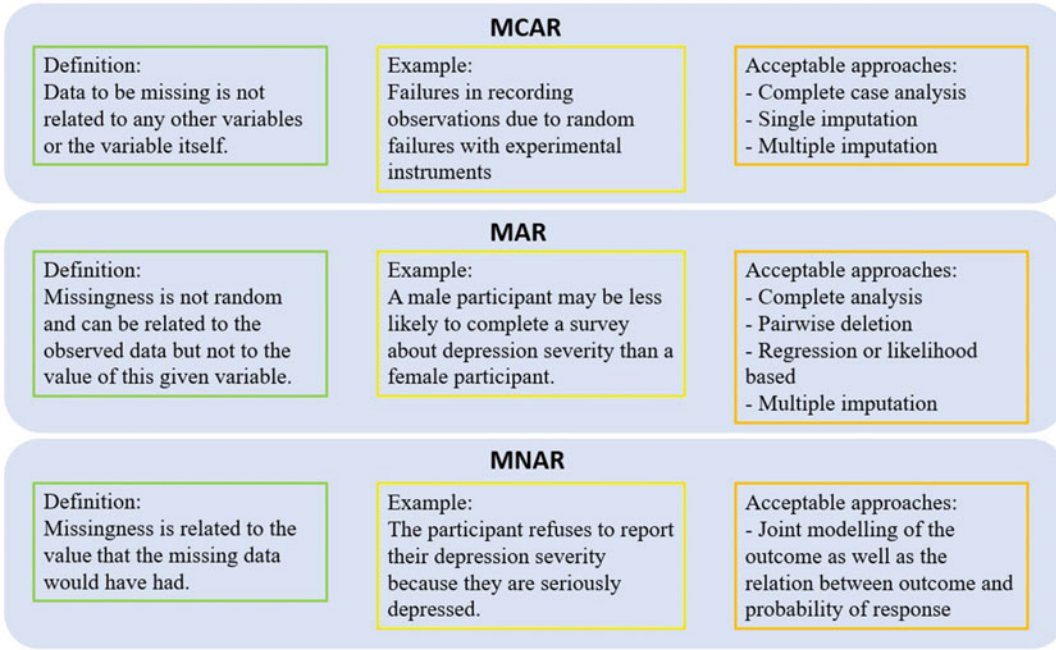


**Fig. 1** Summary on missing mechanisms with definitions, examples, and acceptable approaches for handling the missing values

females. Thus, other observed variables of the participants should be accounted for properly when imputing missing data that are MAR. But MAR is an assumption that is impossible to verify statistically [33] and substantial explorations and analysis are needed. MNAR refers to situations where missingness is related to the value that the missing data would have had. For example, the participant refuses to report their depression severity because they are seriously depressed. In this case, missingness is due to the value itself and no other data can predict this value. Missing data that are MNAR are more problematic as one may lack data from key subgroups which, in turn, may lead to samples that are not representative of the population of interest. The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data and then be incorporated into a more complex one for estimating the missing values [29].

Handling missing data is critical and should be done according to the assumption on the missingness mechanism, as the results might be biased if handled differently. Techniques for handling missing data include the following [29]:

1. Complete case analysis (also known as listwise deletion) to simply omit those cases with the missing data. This approach is suitable for MCAR assumption or when the level of missingness is low in a large dataset.

2. Pairwise deletion allows researchers to use cases with missing values but the variable with missing values will not be included in the analysis. This method is known to be less biased for MCAR or MAR data [29]. The analysis will be deficient if there is a high level of missingness in the data [29].
3. Single imputation means that missing values are replaced by a value defined by a certain rule. Here is a list of possible imputation rules. (1) A simple imputation rule is to substitute the missing value with the mean, median, or mode. (2) A more sophisticated approach uses regression (the missing values are predicted from the other variables using regression). (3) Last observation carried forward or next observation carried backward is for longitudinal data (i.e., repeated measures). If a certain measure is missing, the previous observation or the next observation can be used to impute the current missing values. (4) Maximum likelihood method assumes that the observed data are a sample drawn from a multivariate normal distribution and the missing data are imputed with the maximum likelihood method [34]. (5) K-nearest neighbors method can be used to impute the missing values with the average from the k-nearest neighbors. Single imputation often results in an underestimation of the variability since the unobserved value is analyzed as the known, observed values [35] and some single imputation methods depend on specific rules (e.g., last observation carried forward) rather than missing mechanism assumption which are often unrealistic [36]. Single imputation is often a potentially biased method and should be used with great caution [35–38].
4. Multiple imputation consists in replacing missing values with a set of plausible values which contain the natural variability and uncertainty of the correct values [29]. The multiple imputed values are predicted using the existing data from other variables [39], and then multiple imputed datasets are generated using the set of values. Compared to single imputations, creating multiple imputations accounts for the statistical uncertainty in the imputations. A typical method for multiple imputation is the use of chained equations (MICE) [40]. Multiple imputation operates under the assumption that the missing data are MAR since we use other variables to predict the missing values. Implementing MICE when data are not MAR could result in biased estimates [40]. Multiple imputation has been shown to be a valid method for handling missing data and is considered a good approach for datasets with a large amount of missing data. This method is available for most types of data [31, 37, 38]. Studies comparing software packages for multiple imputations are available [41].

The acceptable handling methods for different missing data mechanisms [35] are summarized in Fig. 1. For MCAR, the methods for handling missing data which give unbiased effects and standard errors are complete case analysis, regression or likelihood-based single imputation methods, and multiple imputation. For MAR assumption, pairwise deletion, regression or likelihood-based single imputation methods, and multiple imputation provide unbiased effects. Under the MNAR assumption, the above methods are no longer suitable. In this case, the appropriate analysis requires the joint modeling of the outcome along with the missing data mechanism [35]. This could be done by asking related questions, e.g., (1) what's the probability of having missing data given the outcome and (2) what's the probability of an outcome in those with missing data? Selection [33] and pattern-mixture models [42] are two example approaches for modeling the above two questions, respectively.

The recommended strategies to overcome barriers caused by missing data would be to first understand the data and the missing mechanism. If the data are simply unavailable, alternative datasets and similar information might be available [28]. Then the imputation method could be selected based on the understanding of the missing values. Since the correctness of the assumptions cannot be definitively validated, it is recommended to perform a sensitivity analysis to evaluate the robustness of the results to the deviations from the assumptions [28].

### 2.3 Other Challenges and General Practices Recommendations

There are other challenges in EHR data. For example, some data may be recorded without specifying units of measurement which makes these data hard to interpret [28]. In this case, an understanding of the data collection process and background knowledge can be helpful in interpreting the data. There might be inconsistencies in data collection and coding across institutions and over time [28]. Some inconsistencies can be easily identified from the data, e.g., a measure was started to be recorded only after a certain time. On the other hand, some inconsistencies may be hard to identify and require an understanding of how data are collected geographically and over time. Last but not least, unstructured text data residing in the EHR causes poor accessibility and other data quality issues such as a lack of objectivity, consistency, or completeness [28]. Data extraction techniques such as natural language processing (NLP) are being used to identify information directly from text notes.

Quality data is the basis for a valid research outcome and whether the quality is enough depends on the purpose of the study. Currently, there are no certain criteria for deciding whether the quality of the data is sufficient, but careful analysis of the data quality should help the researchers decide if the data at hand is useful for the study [28]. Three general practices were

recommended by Feder [20]. The first recommendation is to get familiar with the EHR platform and EHR-based secondary data source. Knowledge of the types of data available, how the data were collected, and who collected it is very useful. It is recommended to have a dictionary that defines all data variables: it should contain the type of data, the range of expected values of each variable, general summary statistics, level of missingness, and subcomponents if available. The second recommendation is to develop a research plan that includes strategies for data quality appraisal and management such as statistical procedures for handling missing data and potential actions if other data quality issues arise (e.g., removal of extreme values, diagnostic code validation). The last recommendation is to promote transparency in reporting data quality including the proportion and type of missing data, other quality limitations, and any subsequent changes made to data values (e.g., variables removed for analysis, imputation methods, variable transformations, creation of new variables). This should enable the reuse of quality data for clinical research. Communications and sharing of the importance of data quality with clinicians are encouraged [28].

## 3 Clinical Coding Systems

In this section, we discuss clinical coding systems, classifications, or terminologies. We first introduce clinical coding systems and explain the motivation behind their existence and usage. This is followed by a discussion of the common attributes that coding systems tend to have, and how this relates to their usage for data analysis. We provide summaries for some of the most commonly used systems in use at the time of writing. Finally, we discuss some of the potential challenges and limitations of clinical coding systems.

### *3.1 Motivation*

Recording clinical data using free text and local terminology incurs major barriers to conducting effective data analysis for health research [43]. Clinical coding systems significantly alleviate this problem, and so are of great usefulness to researchers and analysts when carrying out such work. Medical concepts are naturally described by linguistic terminology and are often associated with a descriptive text. Linguistic data is however loosely structured, and the same underlying medical concept might be expressed differently by different healthcare professionals. Clinical concepts can usually be expressed in a multitude of ways, both due to synonyms in individual terms and simply through different ways of combining and arranging words into a description. Processing large amounts of such data in order to perform modern computer-assisted data analysis, such as training machine learning models, would therefore require the use of natural language processing (NLP) techniques

[44]. Furthermore, when considering medical data from many countries, one would need to consider all the possible languages that medical records might be written in.

Instead of mapping clinical concepts into the highly complex realm of natural language, clinical coding systems seek to provide an unambiguous mapping from a given clinical concept to a unique encoding in a principled fashion. This makes it significantly easier to employ modern large-scale data analysis techniques on clinical data. For example, if one were interested in studying the prevalence of chronic fatigue, instead of having to attempt to exhaustively match records containing every conceivable way to express this linguistically, one would only need to identify which clinical codes are associated with the relevant clinical concepts and select records containing those codes.

### 3.2 Common Characteristics

Clinical coding systems can vary significantly in their descriptive scope, depending on their intended usage. The DSM-5 [45], for instance, limits its scope entirely to psychiatric diagnoses, while SNOMED-CT [46, 47] seeks to be as comprehensive as possible, including concept codes relating to, for example, body structure, physical objects, and environment. Both of these coding schemes describe concepts relevant at the level of individual patients, though codes can exist for broader or more fine-grained scopes such as public health or microbiology.

Typically, clinical coding schemes are arranged hierarchically, as this reflects the categorical relationship between clinical concepts well while also providing an intuitive means to find relevant concepts. This hierarchical structuring can be reflected in the identifiers used to encode clinical concepts, further aiding in their comprehension. In the ICD scheme [48], for example, codes begin with a character that identifies the relevant chapter in the ICD manual, and subsequent characters provide identification of finer and finer degrees of specification.

Another property of clinical coding systems that can be useful to classify is whether it is *compositional* or *enumerative* [49, Chapter 22]. In a compositional scheme, concepts can be encoded by combining more basic conceptual units together. This reduces the burden to specify large enough lists of distinct concepts to comprehensively cover all necessary clinical concepts required by scheme designers. This is in contrast to enumerative systems, which instead aim to achieve completeness by having a unique identifier for every concept within the scope of the scheme.

Clinical coding schemes can encode many kinds of relationships between concepts that are more specific than the simple parent-child relationship in basic hierarchies. These reflect the more nuanced kinds of relationships present in clinical concepts. Coiera [49, Chapter 22] outlines three main kinds of conceptual relationships: `Part-Whole`, `Is-A`, and `Causal`. `Part-Whole`

relationships are useful when a concept contains constituent parts which are also concepts, e.g., the eyes are a part of the face which is a part of the head which is a part of the body. This relationship is generally most useful for describing physical assemblages. `Is-A` relationships are perhaps the most common and indicate basic categorical similarities, such as Arterial Blood Specimen `Is-A` Blood Specimen `Is-A` Specimen. Finally, Causal relationships are used to indicate events or effects that arise as the result of another, or that cause another.

Hierarchical schemes may also introduce multiple axes upon which to expand concepts (essentially multiple hierarchies). In this way, elements belonging to a particular place in the hierarchy of one axis may also appear in the hierarchy of a different axis. This often involves a concept having multiple relationships of different types to a number of different concepts, i.e., a concept may have an `Is-A` relationship and a `Causal` relationship with two different concepts.

These are all useful features in the context of data science. Hierarchical structures allow for users of data to select as coarse or as fine-grained concepts as are relevant to their specific analyses. The defined relationships between concepts can be exploited in order to identify groups of relevant codes. Furthermore, some coding schemes, such as SNOMED-CT, may encode useful concepts beyond clinical events or concepts, such as whether patients have consented for research data usage, which can be useful, for example, in screening population members who are unsuitable for research cohorts, etc.

### 3.3 Notable Coding Systems

Here we provide summaries of commonly used coding systems that are likely to be encountered when performing analysis on EHR data. However, this is by no-means an exhaustive list. Many more are in use, and some datasets or corpora might use their own coding systems. In these cases, the data provider will usually specify mappings to more common systems such as ICD or SNOMED-CT. For example, in the case of the Clinical Practice Research Datalink (CPRD) [50], unique codes are provided for medical terms with mappings to Read Codes (a now largely legacy coding system in the United Kingdom), and unique treatment codes with links to the NHS Dictionary of Medicines and Devices (dm+d) [51] and the British National Formulary (BNF) [52], which provide codes relating specifically to medical products and prescribing.

#### 3.3.1 SNOMED-CT

SNOMED-CT (Systematized NOmenclature of MEDicine-Clinical Terms) [46], maintained by SNOMED International, is a clinical coding scheme designed to be highly comprehensive and computer-processable. It is in wide usage around the world, in particular in the United Kingdom. SNOMED-CT supersedes the older SNOMED and SNOMED-RT systems. It is a hierarchical, compositional coding scheme, including specified relationships

**Table 1**
**The top-level hierarchical categories in the SNOMED-CT system**

| Hierarchy |
|---|
| Body structure |
| Clinical finding |
| Event |
| Observable entity |
| Organism |
| Pharmaceutical/biologic product |
| Physical object |
| Procedure |
| Qualifier value |
| Situation with explicit context |
| Social context |
| Substance |

between related concepts. It provides good linkage with ICD to allow for easy data sharing. There are 15 primary hierarchical categories in SNOMED-CT, to which all other concepts belong. A concept in SNOMED-CT is comprised of several elements. The primary identifying element is the Concept ID, which is a unique numerical identifier for the clinical concept. This is accompanied by a textual description of the concept. There are specified Relationships to other related concepts, and Reference Sets which provide groupings of concepts. SNOMED-CT codes are hierarchical and linked via `Is-A` relationships. Table 1 presents the top-level concepts of SNOMED-CT.

#### 3.3.2 ICD

The ICD (International Classification of Diseases) [48] is a coding system created by the World Health Organization (WHO). While the ICD is currently in its 11th revision (ICD-11) [53], ICD-10 is still more commonly used at the time of writing, and the widespread adoption of ICD-11 will likely take more time. The ICD system is a multi-axis hierarchical coding system, assigning an alphanumeric code to each concept. Each code is procedurally derived from its concept's location in the hierarchy, aiding in comprehension. The first character letter in an ICD code associates it with a specific chapter in the ICD manual (see Table 2 for the different chapters of ICD-10). The following three characters locate the concept within the chapter and range from A00 to Z99. For more detail, each category can be further subdivided

**Table 2**
**The chapters of ICD-10**

| Number | Chapter name |
|---|---|
| I | Certain infectious and parasitic diseases |
| II | Neoplasms |
| III | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism |
| IV | Endocrine, nutritional, and metabolic diseases |
| V | Mental and behavioral disorders |
| VI | Diseases of the nervous system |
| VII | Diseases of the eye and adnexa |
| VIII | Diseases of the ear and mastoid process |
| IX | Diseases of the circulatory system |
| X | Diseases of the respiratory system |
| XI | Diseases of the digestive system |
| XII | Diseases of the skin and subcutaneous tissue |
| XIII | Diseases of the musculoskeletal system and connective tissue |
| XIV | Diseases of the genitourinary system |
| XV | Pregnancy, childbirth, and the puerperium |
| XVI | Certain conditions originating in the perinatal period |
| XVII | Congenital malformations, deformations, and chromosomal abnormalities |
| XVIII | Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified |
| XIX | Injury, poisoning, and certain other consequences of external causes |
| XX | External causes of morbidity and mortality |
| XXI | Factors influencing health status and contact with health services |
| XXII | Codes for special purposes |

with up to three additional numeric characters. Table 3 shows multiple sclerosis as it appears in ICD-11 as an example of this hierarchical coding structure. The ICD system is intended to be limited in scope to disease diagnosis-related concepts; however, the WHO maintains additional systems to cover concepts outside of this scope. The ICF (International Classification of Functioning, Disability and Health), for instance, focuses on a patient's capacity to live and function and includes concepts relating to body functions, bodily structures, activities, participation, and environmental factors. Furthermore, various modifications of the ICD system exist to expand upon its capabilities for use in clinical settings, such as the ICD-10-CM in the United States and the ICD-10-CA in Canada.

**Table 3**
**The hierarchical structure of multiple sclerosis within the ICD-11**

| 1. ICD-11 for Mortality and Morbidity Statistics |
|---|
| • **08** - Diseases of the nervous system |
| – Multiple sclerosis or other white matter disorders |
| ∗**8A40** - Multiple sclerosis |
| ·**8A40.0** - Relapsing-remitting multiple sclerosis |
| ·**8A40.1** - Primary progressive multiple sclerosis |
| ·**8A40.2** - Secondary progressive multiple sclerosis |
| ·**8A40.Y** - Other specified multiple sclerosis |
| ·**8A40.Z** - Multiple sclerosis, unspecified |

### 3.3.3 UMLS

"The Unified Medical Language System (UMLS) is something like the Rosetta Stone of international terminologies"—Coeira [49, Chapter 23]

The UMLS [54] is intended to provide a means to relate coding systems to each other. It achieves this with three knowledge sources: the Metathesaurus, a semantic network, and the SPECIALIST Lexicon. The Metathesaurus is a nonhierarchical controlled vocabulary of terms organized by concept and provides the synonyms of concepts in different coding systems and is the primary way in which translation between systems is supported. Controlled vocabularies from hundreds of coding systems are represented in the Metathesaurus, and its entries are regularly updated. A complete list of all the supported controlled vocabularies is available in the UMLS Metathesaurus Vocabulary Documentation on the official website.[1] The Metathesaurus specifies defining attributes of concepts, and relationships between concepts, including `Is-A`, `Part-Whole`, and `Causal` relationship types. The semantic network provides the semantic types and relationships that concepts are permitted to inherit from. The primary semantic relationship is the hierarchical `Is-A` relationship, although there are five primary nonhierarchical relationship types: "physically related to," "spatially related to," "temporally related to," "functionally related to," and "conceptually related to." The SPECIALIST Lexicon is intended to assist computer applications in interpreting free-text fields. It encodes syntactic, morphological, and orthographic information, including common spelling variants. In practice, most users of the UMLS do so indirectly through

[1] https://www.nlm.nih.gov/research/umls/index.html.

tools that rely on the UMLS, such as PubMed[2] and other clinical software systems such as EHR software and analysis pipelines. The most common uses are for extracting clinical terminologies from text and translating between coding systems [55].

#### 3.3.4 Read Codes

Read Codes [56, 57] were used exclusively by the United Kingdom until 2018, when they were replaced by SNOMED-CT. Read Codes are organized hierarchically; however, the identifiers themselves do not indicate where in a hierarchy the concept belongs as they do in ICD. Version 3 (CTV-3) is the most recent version of Read Codes, and introduced compositionality to the system, while becoming less strictly hierarchical. Read Codes were intended to provide digital operability in primary care settings, but are no longer used in primary care in England (though they are still in use in Scotland at the time of writing and may be used in secondary care in England). Read Codes map well to ICD concepts. The Read Codes Drug and Appliance Dictionary is an extension of the Read Codes system to include pharmacological products, foods, and medical appliances for use in EHR software and prescribing systems.

### 3.4 Challenges and Limitations

The usefulness of clinical coding schemes is dependent upon their usage by healthcare professionals being thorough and appropriate. Improper usage of coding systems can occur, contributing to data quality issues such as incompleteness, inconsistency, and inaccuracy [58]. Further challenges can arise for researchers where data may contain multiple coding systems; this can happen if the data is collected from multiple different sources where different coding systems are in use, or if the period of data collection covers a change in the preferred coding system, such as the change from CVT-3 to SNOMED-CT in the United Kingdom. In these cases, the researcher must ensure that they consider relevant concepts from each different scheme or implement a mapping from one scheme to another. Most coding schemes provide good mapping support to ICD codes, and the UMLS coding system is designed to provide a means of translating between different schemes. Additionally, some sources of data may provide their own coding schemes that are not in usage (and thus not documented) elsewhere.

## 4 Protection and Governance of EHR Data

In this section, we will explore the focal points of data protection and governance analyzing the most recent jurisdictional background and its implication in real-world healthcare applications. In Subheading 4.1, we introduce the main legislative body and its

[2] https://pubmed.ncbi.nlm.nih.gov/.

core definitions in data protection. Then, Subheading 4.2 describes in a more technical way how data analysis can be conducted in a privacy-preserving manner.

### 4.1 Data Protection in a Nutshell

The explosive evolution of digital technologies and our ability to collect, store, and elaborate data is dramatically changing how we should consider privacy and data protection; particularly, the advent of artificial intelligence (AI) and advanced mathematical modeling tools made it necessary to reform the national and international data protection and governance rules to better protect people who generated such data and give them more control on what can be done with it. Although it is worth mentioning valuable independent contributions to the healthcare data protection guidelines like the Goldacre Review [59, 60], we will focus mainly on the most recent and structured action published at international level in terms of data protection and governance, the European General Data Protection regulation, or GDPR [17].

The GDPR was published by the European Commission in 2016 to set the guidelines that all member states must apply in their national legislation in terms of data protection. Although its legal validity is limited to the members of the European Economic Area (EEA), its effects expanded also to European Union (EU) candidate countries and the United Kingdom which embraced the new GDPR regulation through the UK GDPR [18] and maintained it part living of the legislation even after renouncing to the EU membership. It is worth mentioning that the effects of GDPR are not limited to the data management and governance executed within the countries that embrace the regulation, but is strictly related to the persons to whom the data belong; this means that the GDPR guidelines must be followed by any entity worldwide when dealing with data belonging to individuals from countries where the GDPR applies. GDPR defines as *personal data* any single information that is relatable to a person; in Box 1 we enumerate the three main agents required in any endeavor involving personal data management.

To contextualize these concepts in an healthcare scenario, if a non-European controller (e.g., an Australian hospital) aims at collecting, storing, or elaborating healthcare data from an individual protected by the GDPR or equivalent legislation for an international multicenter clinical trial, they still must respect all dictations of GDPR on that data specifically.

The GDPR reads: *personal data processing should be designed to serve mankind and the right to the protection of such data is not an absolute right, but must be considered in relation to its function in society.* Let's then consider this from the two angles of data governance and operation, and its purpose in the AI era.

**Box 1: Basic agents recognized by GDPR**

| | |
|---|---|
| **Data subject** | Individual(s) to whom the personal data belongs. |
| **Controller** | Individual(s) or institution(s) responsible for implementing appropriate technical and organizational measures to ensure and to be able to demonstrate that processing is performed in accordance with the GDPR. |
| **Processor** | Individual(s) or institution(s) responsible for using, manipulating, and leveraging personal data for the goals defined by the controller and agreed upon by the data subject. |

#### *4.1.1 Governance and Operation*

One of the main dictations of GDPR is that data should be as anonymized (or, de-identified) and minimal as possible for a given application. This means that the data controller shall specify in details which data will be needed and why and collect only this required data, possibly in an anonymous way. Moreover, the data should be stored as long as the application requires it but not longer unless authorized by the data subject. This process should minimize as much as possible the identifiability of individuals, especially in those cases in which the content of data carries very sensitive information like health status, religious faith, political affiliation, and similar. Indeed, one of the main reasons why the use of free-text clinical notes in natural language processing (NLP) applications carries additional complications is that information that could identify individuals are often expressed in a nonstructured way in text (e.g., a specific reference to a person's habits, rare diseases, physical aspect, etc.) [61]. A similar issue arises with imaging applications, where the content of the imaging medical examination could contain personal information of its owner (e.g., the name written on an X-ray printing).

With a closer focus to EHR in a common tabular structure, identification of individuals can go beyond their names and unique identifiers. If the combination of other information can lead to their identification (e.g., the address, the sex, physical characteristics, profession, etc.), then the EHR is not technically anonymized. A step forward is the pseudo-anonymization, a process where the identifiable information fields are replaced with artificially created alternatives that encode or encrypt these information without direct disclosure. It is important to note that albeit this approach is valid in healthcare applications, it still allows a post hoc reconstruction of the identifiable data and should be implemented

carefully. Note that, in the specific case of brain images, the medical image may in principle allow reidentification of the patient (for instance, mainly through recognition of facial features such as the nose). For this reason, "defacing" (a procedure that modifies the image to remove facial features while preserving the content of the brain) is increasingly used. According to the Health Insurance Portability and Accountability Act of 1996 (HIPAA)[3] issued by the US Department of Health and Human Services, 18 elements have to be deleted for an electronic health record to be considered de-identified; these include names, geographic subdivision smaller than a State, all elements of dates (except year) for dates directly related to an individual, telephone numbers, social security numbers, and license numbers. This practice can be exported internationally and used as a rule of thumb to ensure appropriate anonymization in all healthcare-related applications.

With respect to the many stages that comprise the analysis and elaboration of healthcare data, data protection can be handled in different and more flexible ways. Assuming a high level of internal protection of healthcare institutions (e.g., firewalls and encrypted servers), as long as the data remains within the institution secured information system, the majority of threats can be blocked and mitigated at an institutional level. Examples of threats are malicious access to and modification of data with the objective of compromising individual's health or disrupting the operation of the hospital itself. The main exposure happens in case the data need to be transferred to another institution to carry out the required analyses. In this rather common case, the anonymization (or pseudo-anonymization) process should be carefully applied and data should never reside in a non-secured storage device or communication channel. To prevent this exposure to happen but at the same time to leave the possibility of leveraging the collected data for the purpose of AI applications and statistical studies, the federated learning methodology has been developed in recent years. This will be described further in Subheading 4.2.

The data subject has the right to get its own data deleted from the controller when, for example, the accuracy of the data is contested by the data subject, or when the controller no longer needs the data for its purposes. Similarly, the data subject has also the right to receive their personal data from the controller in a commonly used and machine-readable format and have then the right to transfer such data to another controller, when technically feasible, in a direct way. These aspects introduce operational constraints in EHR management as they require to be stored in an identifiable way (so as to allow its post hoc management, deletion, or

[3] https://www.hhs.gov/hipaa/index.html.

modification) but to be elaborated in an non-identifiable manner to ensure that at any point of the data elaboration, the identification of the patients is impossible or as minimal as required for the elaboration itself. A corner case would be when a patient revokes the right of the controller to handle their data and its anonymized version is in use; from an operational point of view, this could cause the need for re-execution of the data extraction and elaboration.

#### 4.1.2 The Purpose of EHR in the Era of AI

The main conundrum here is whether a specific use of healthcare data is functional to a societal benefit, which is a very difficult problem given its highly subjective interpretation. Indeed, as we continue producing beneficial applications, the opportunities to develop malevolent ones increase. Hostile actors may use private healthcare data and AI for personal profits, policy control, and other malicious cases. The availability of new tools suddenly sheds light on problems we didn't know we had and this is happening with AI and its application to healthcare. Machine learning and deep learning are by far the most successful technologies that are changing how we conceive data value and the importance of its quality [62], and when it comes to these computing tools, the more data, the better, but not only that; for each application, the data collected and elaborated should be as representative as possible of the learning task, which is a rather challenging issue considering the amount of human intervention in clinical data collection (especially in free-text annotations) and inherent biases in the data distribution over the available population. Current regulations are imposed to the data controllers to clearly communicate and have the explicit agreement of the data subject for any use they may do with it, and this is a fundamental protection of each individual's right to choose when and where their data can be used. This becomes particularly stressed in healthcare scenarios where misuse and abuses of patients' data can result in unethical advantage and/or enrichment of the institutions or individuals capable of making the most out of such abundant data.

Ethical approvals for the use of clinical datasets are usually granted by the hospitals' ethic committee, through detailed processes that every study has to undertake in its design phase. However, with increased focus on the use of AI technologies in medicine, the challenge becomes to contextualise within these ethics frameworks new technologies, the potential they carry, and the risks they may represent. Therefore, an integrated approach is needed between clinical experts, and AI/ML specialists to give more transparency, cohesion, and consistency to the use of data in health research.

### 4.2 Privacy-Preserving EHR Data Analysis

The training of any kind of AI-based predictive model requires as much data as possible, and given the nature of clinical data (costly and with high human intervention), it is often the case that a single healthcare institution is not enough to produce the data needed for the creation of a predictive model. This is particularly true in those cases in which the distribution of the population of patients within the hospital is not representative of the general population or at least of the possible population of patients for which those predictive models will be used.

The most straightforward practice to overcome this limitation consists in gathering data from multiple institutions in one single center and pre-process the data so as to integrate everything in one single training dataset. This allows the unification of the contribution of all healthcare institutions and therefore a more comprehensive, heterogeneous, and representative training dataset. Transferring clinical data from one hospital to another is a procedure that brings many privacy- and security-related problems, including the proper anonymization, or pseudo-anonymization, of clinical records and the encryption of the data en route to another institution.

The technical difficulties here dominate over the potential of a scalable, efficient, and secure data science pipeline that properly uses EHR to extract new knowledge and train predictive models.

One of the most brilliant solutions to solve these problems was initially proposed by Google with the federated learning methodology [63]. According to this approach designed primarily for deep neural networks, instead of transferring the data between institutions and collect everything in one unique dataset, a more efficient choice is to send the models to be trained to every institution that participates in the federation and, once one or more training steps are executed, gather the trained models in one central computing node (which can be one of the institutions) and compile the trained models in one comprehensive unique solution that represents the common knowledge produced.

Federated learning was designed for a task very different from clinical applications, i.e., the automatic completion of smartphones' keyboard, but its principles can be translated to the healthcare environment very effectively. The main benefits are that clinical data will never leave the owner's secured information system and anonymization and encryption of the data itself are not major problems. Moreover, the ability to involve the contribution of multiple centers for one training process requires a software infrastructure that can be utilized many more times for learning tasks.

### 4.3 Challenges Ahead

In the context of federated learning for EHR analysis, we find many challenges to be addressed in terms of both data quality and governance and learning methodologies. Here are listed some of the most relevant:

1. Not having direct access to other institutions' data makes it harder to assess the quality, consistency, and completeness of the datasets. This mandates additional care to the learning strategies as the representativeness of data must be preserved and phenomena like the *catastrophic forgetting* [64] produced by a large amount of data should be prevented.
2. Even assuming a good enough data quality in terms of completeness, correctness, and standards used, the distribution of data in independent datasets can be very different, posing additional learning challenges in the creation of a reliable and fair predictive model. This phenomenon is also known as the *non-IID*, or *non-independent and identically distributed*, data, and it is a very active research field [65].
3. Regardless of the immobility of data in healthcare information systems, the predictive models still have to travel between institutions, and this allows the possibility of data reconstruction through inverse gradient strategies [66], and the predictive model alteration (or *poisoning*) [67, 68] to induce it to behave in a malicious way; this transfers the security problems from the data to the machine learning models themselves and must be properly dealt both at a network level (with encrypted connections) and at a model level to mitigate communication bottlenecks, poisoning, backdoor, and inference-based attacks [69].

## 5 Conclusion

Increasing interest and opportunities for various research purposes were attracted by the rapidly growing number of EHRs. To draw valid and reliable research findings, data quality is paramount. In this chapter, we first introduced the definition of data quality, the reported components, and the concerns raised with poor data quality. Various aspects of data quality components and challenges were explored, such as data accuracy and data completeness. General practices for data quality analysis were recommended at the end of the data quality section.

We then introduced the concepts of a clinical coding system and discuss their potential challenges and limitations. We described the common characteristics of coding systems and then presented some of the most common ones: SNOMED-CT, ICD, UMLS, and Read Codes.

Finally, we navigated the main concepts of data governance and protection in healthcare settings. National and international regulations are put in place to define baseline principles to ensure the most appropriate treatment, storage, and final utilization of personal data, including healthcare information. From an operational perspective, there are numerous challenges to face, e.g., the

anonymization, or pseudo-anonymization, of patients' data and its proper privacy-preserving analysis for business and clinical purposes. This is particularly important in machine learning applications where a large amount of data is required and data sharing between hospitals is not a viable and secure solution. To produce a truly privacy-preserving approach for machine learning applications, federated learning is today the most effective and promising deployable methodology.

## Acknowledgements

This research was funded/supported by the National Institute for Health and Care Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London and/or the NIHR Clinical Research Facility. WW and VC acknowledge the financial support from the Health Foundation. GHH and VC were supported by Innovation Scholars Big Data and AI Training MR/V038664/1 MRC funded by The Medical Research Council (MRC) at King's College London. The views expressed are those of the author(s) and not necessarily those of the Health Foundation, the MRC, the NHS, the NIHR, or the Department of Health and Social Care.
The authors are grateful to Prof. Paolo Missier for reviewing this chapter and providing useful insight.

### References

1. CPRD (n.d.) Clinical practice research datalink. https://cprd.com/
2. QResearch (n.d.) QResearch. https://www.qresearch.org/
3. ResearchOne (n.d.) Transforming data into knowledge. http://www.researchone.org/
4. Alliance UHDR (2020) Hdruk innovation gateway — homepage. https://www.healthdatagateway.org/
5. Verheij R, van der Zee J (2018) Collecting information in general practice: "just by pressing a single button"? Morbidity, Performance and Quality in Primary Care pp 265–272. https://doi.org/10.1201/9781315383248-36
6. Nivel (n.d.) Nivel primary care database. https://www.nivel.nl/en/nivel-zorgregistraties-eerste-lijn/nivel-primary-care-database
7. Schweikardt C, Verheij RA, Donker GA, Coppieters Y (2016) The historical development of the dutch sentinel general practice network from a paper-based into a digital primary care monitoring system. J Public Health (Germany) 24:545–562. https://doi.org/10.1007/S10389-016-0753-4/TABLES/3. https://link.springer.com/article/10.1007/s10389-016-0753-4
8. Bartholomeeusen S, Kim CY, Mertens R, Faes C, Buntinx F (2005) The denominator in general practice, a new approach from the intego database. Fam Pract 22:442–447. https://doi.org/10.1093/FAMPRA/CMI054. https://academic.oup.com/fampra/article/22/4/442/662730
9. SNDS (n.d.) Système national des données de santé. https://www.bordeauxpharmacoepi.eu/en/snds-presentation/
10. Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P, Moore N (2017) The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology. Pharmacoepidemiol Drug Saf 26(8):954–962
11. Daniel C, Salamanca E (2020) Hospital Databases: AP-HP data warehouse. In: Nordlinger B, Villani C, Rus D (eds)

Healthcare and artificial intelligence. Springer, Berlin, pp 57–67
12. Ludvigsson JF, Almqvist C, Bonamy AKE, Ljung R, Michaëlsson K, Neovius M, Stephansson O, Ye W (2016) Registers of the Swedish total population and their use in medical research. Eur J Epidemiol 31(2):125–136
13. Serda M (2013) Synteza i aktywność biologiczna nowych analogów tiosemikarbazonowych chelatorów żelaza
14. Gliklich RE, Dreyer NA, Leavy MB (2014) Registries for evaluating patient outcomes. AHRQ Publication 1:669. https://www.ncbi.nlm.nih.gov/books/NBK208616/
15. Fleurence RL, Beal AC, Sheridan SE, Johnson LB, Selby JV (2017) Patient-powered research networks aim to improve patient care and health research. Health Aff 33(7):1212–1219. https://doi.org/10.1377/HLTHAFF.2014.0113
16. CTSA (n.d.) CTSA Central. http://www.ctsacentral.org/
17. GDPR (2016) EU General Data Protection Regulation. http://data.europa.eu/eli/reg/2016/679/oj
18. UK GDPR (2018) UK General Data Protection Regulation Updated for Brexit — UK GDPR. https://uk-gdpr.org/
19. Fundation TM (2006) Background issues on data quality. In: The connecting for health common framework https://bok.ahima.org/PdfView?oid=63654
20. Feder SL (2018) Data quality in electronic health records research: quality domains and assessment methods. West J Nurs Res 40(5):753–766. https://doi.org/10.1177/0193945916689084
21. Chan KS, Fowles JB, Weiner JP (2010) Review: Electronic health records and the reliability and validity of quality measures: A review of the literature. Med Care Res Rev 67(5):503–527. https://doi.org/10.1177/1077558709359007
22. Kahn M, Raebel M, Glanz J, Riedlinger K, Steiner J (2012) A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care 50(Suppl):S21–9. https://doi.org/10.1097/MLR.0b013e318257dd67
23. Wand Y, Wang RY (1996) Anchoring data quality dimensions in ontological foundations. Commun ACM 39(11):86–95. https://doi.org/10.1145/240455.240479
24. Weiskopf NG, Weng C (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 20(1):144–151. https://doi.org/10.1136/amiajnl-2011-000681. https://academic.oup.com/jamia/article-pdf/20/1/144/9517051/20-1-144.pdf
25. Ahmad F, Rasmussen L, Persell S, Richardson J, Liss D, Kenly P, Chung I, French D, Walunas T, Schriever A, Kho A (2019) Challenges to electronic clinical quality measurement using third-party platforms in primary care practices: The healthy hearts in the heartland experience. JAMIA Open 2(4):423–428. https://doi.org/10.1093/jamiaopen/ooz038
26. Tse J, You W (2011) How accurate is the electronic health record?—a pilot study evaluating information accuracy in a primary care setting. Stud Health Technol Inform 168:158–64
27. Ozair F, Nayer J, Sharma A, Aggarwal P (2015) Ethical issues in electronic health records: A general overview. Perspect Clin Res 6:73–6. https://doi.org/10.4103/2229-3485.153997
28. Bayley K, Belnap T, Savitz L, Masica A, Shah N, Fleming N (2013) Challenges in using electronic health record data for CER: Experience of 4 learning organizations and solutions applied. Med Care 51:S80-S86. https://doi.org/10.1097/MLR.0b013e31829b1d48
29. Hyun K (2013) The prevention and handling of the missing data. Korean J Anesthesiol 64(5):402–406. https://doi.org/10.4097/kjae.2013.64.5.402. http://ekja.org/journal/view.php?number=7569
30. Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592. http://www.jstor.org/stable/2335739
31. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338. https://doi.org/10.1136/bmj.b2393. https://www.bmj.com/content/338/bmj.b2393. https://www.bmj.com/content
32. Smith WG (2008) Does gender influence online survey participation? A record-linkage analysis of university faculty online survey response behavior. Online Submission
33. Little R, Rubin D (2002) Statistical analysis with missing data. In: Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, London. http://books.google.com/books?id=aYPwAAAAMAAJ
34. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via

the em algorithm. J R Stat Soc Ser B Methodol 39(1):1–38. http://www.jstor.org/stable/2984875

35. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P (2013) Strategies for dealing with missing data in clinical trials: from design to analysis. Yale J Biol Med 86(3):343–358. https://europepmc.org/articles/PMC3767219
36. Jakobsen JC, Gluud C, Wetterslev J, Winkel P (2017) When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. BMC Med Res Methodol 17(1):1–10
37. Zhang Y, Flórez ID, Lozano LEC, Aloweni FAB, Kennedy SA, Li A, Craigie SM, Zhang S, Agarwal A, Lopes LC, Devji T, Wiercioch W, Riva JJ, Wang M, Jin X, Fei Y, Alexander PE, Morgano GP, Zhang Y, Carrasco-Labra A, Kahale LA, Akl EA, Schünemann HJ, Thabane L, Guyatt GH (2017) A systematic survey on reporting and methods for handling missing participant data for continuous outcomes in randomized controlled trials. J Clin Epidemiol 88:57–66
38. Jørgensen AW, Lundstrøm LH, Wetterslev J, Astrup A, Gøtzsche PC (2014) Comparison of results from different imputation techniques for missing data from an anti-obesity drug trial. PLoS One 9(11):1–7. https://doi.org/10.1371/journal.pone.0111964
39. Sinharay S, Stern H, Russell D (2001) The use of multiple imputation for the analysis of missing data. Psychol Methods 6:317–29. https://doi.org/10.1037/1082-989X.6.4.317
40. Azur M, Stuart E, Frangakis C, Leaf P (2011) Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 20:40–9. https://doi.org/10.1002/mpr.329
41. Horton NJ, Lipsitz SR (2001) Multiple imputation in practice: comparison of software packages for regression models with missing variables. Am Stat 55(3):244–254. http://www.jstor.org/stable/2685809
42. Little RJA, Wang Y (1996) Pattern-mixture models for multivariate incomplete data with covariates. Biometrics 52(1):98–111. http://www.jstor.org/stable/2533148
43. Elkin PL, Trusko BE, Koppel R, Speroff T, Mohrer D, Sakji S, Gurewitz I, Tuttle M, Brown SH (2010) Secondary use of clinical data. Stud Health Technol Inform 155:14–29
44. Koleck TA, Dreisbach C, Bourne PE, Bakken S (2019) Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 26(4):364–379. https://doi.org/10.1093/jamia/ocy173
45. Association AP, Association AP (eds) (2013) Diagnostic and statistical manual of mental disorders: DSM-5, 5th edn. American Psychiatric Association, Arlington, VA, oCLC:830807378
46. SNOMED International (2022) SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/index.html. publisher: U.-S. National Library of Medicine
47. Lee D, de Keizer N, Lau F, Cornet R (2014) Literature review of SNOMED CT use. J Am Med Inform Assoc 21(e1):e11–e19. https://doi.org/10.1136/amiajnl-2013-001636
48. World Health Organisation (2022) International Classification of Diseases (ICD). https://www.who.int/standards/classifications/classification-of-diseases
49. Coiera E (2015) Guide to health informatics. CRC Press, Boca Raton. google-Books-ID: 1ngZBwAAQBAJ
50. Medicines and Healthcare products Regulatory Agency (2022) Clinical Practice Research Datalink | CPRD. https://www.cprd.com
51. NHS (2022) Dictionary of medicines and devices (dm+d) — nhsbsa. https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/dictionary-medicines-and-devices-dmd
52. Committee JF (2022) BNF (British national formulary) — nice. https://bnf.nice.org.uk/
53. Organisation WH (ed) (2019) International statistical classification of diseases and related health problems, 11th edn. World Health Organization, New York. https://icd.who.int/
54. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32 (Database issue):D267–D270. https://doi.org/10.1093/nar/gkh061. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308795/
55. Amos L, Anderson D, Brody S, Ripple A, Humphreys BL (2020) UMLS users and uses: a current overview. J Am Med Inform Assoc 27(10):1606–1611. https://doi.org/10.1093/jamia/ocaa084
56. NHS (2020) Read Codes. https://digital.nhs.uk/services/terminology-and-classifications/read-codes
57. N B (1994) What are the Read Codes? Health Libr Rev 11(3):177–182. https://doi.org/10.1046/j.1365-2532.1994.1130177.x. https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2532.1994.1130177.x
58. Botsis T, Hartvigsen G, Chen F, Weng C (2010) Secondary use of EHR: data quality

issues and informatics opportunities. Summit on Translational Bioinformatics 2010:1–5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041534/

59. Ben Goldacre ea (2022a) Better, broader, safer: using health data for research and analysis—gov.uk. https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis
60. Ben Goldacre ea (2022b) Home — goldacre review. https://www.goldacrereview.org/
61. Pan X, Zhang M, Ji S, Yang M (2020) Privacy risks of general-purpose language models. Proceedings—IEEE Symposium on Security and Privacy 2020(May):1314–1331. https://doi.org/10.1109/SP40000.2020.00095
62. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J (2019) A guide to deep learning in healthcare. Nat Med 25(1):24–29. https://doi.org/10.1038/s41591-018-0316-z. https://www.nature.com/articles/s41591-018-0316-z
63. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAy (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Singh A, Zhu J (eds) Proceedings of the 20th international conference on artificial intelligence and statistics, PMLR, Proceedings of Machine Learning Research, vol 54, pp 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html
64. McCloskey M, Cohen NJ (1989) Catastrophic interference in connectionist networks: the sequential learning problem. In: Bower GH (ed) Psychology of learning and motivation, vol 24, Academic Press, New York, pp 109–165. https://doi.org/10.1016/S0079-7421(08)60536-8. https://www.sciencedirect.com/science/article/pii/S0079742108605368
65. Zhu H, Xu J, Liu S, Jin Y (2021) Federated learning on non-IID data: A survey. Neurocomputing 465:371–390. https://doi.org/10.1016/j.neucom.2021.07.098, 2106.06843
66. Geiping J, Bauermeister H, Dröge H, Moeller M (2020) Inverting gradients—how easy is it to break privacy in federated learning? In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates Inc, New York, pp 16937–16947. https://proceedings.neurips.cc/paper/2020/file/c4ede56bbd98819ae6112b20ac6bf145-Paper.pdf
67. Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V (2020) How to backdoor federated learning. In: Chiappa S, Calandra R (eds) Proceedings of the twenty third international conference on artificial intelligence and statistics, PMLR, proceedings of machine learning research, vol 108, pp 2938–2948. https://proceedings.mlr.press/v108/bagdasaryan20a.html
68. Lyu L, Yu H, Zhao J, Yang Q (2020) Threats to federated learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12500 LNCS:3–16. https://doi.org/10.1007/978-3-030-63076-8_1. https://arxiv.org/abs/2003.02133v1, 2003.02133
69. Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G (2021) A survey on security and privacy of federated learning. Futur Gener Comput Syst 115:619–640. https://doi.org/10.1016/j.future.2020.10.007. https://www.sciencedirect.com/science/article/pii/S0167739X20329848