

Characterisation of Data Quality in Electronic Healthcare Records

Sheena Dungey, Natalia Beloff, Rachael Williams, Tim Williams, Shivani Puri, and A. Rosemary Tate

1 Introduction

The use of electronic healthcare systems for recording patient treatment history is well established across the UK healthcare sector, the potential benefits of using such systems being numerous. Within the primary care setting, electronic healthcare records (EHR) can provide a near complete picture of patient care over time. This not only affords the opportunity to improve patient care directly through effective monitoring and identification of care requirements but also offers a unique platform for both clinical and service-model research [1] essential to the longer term development of the health service. The potential for using routinely collected patient records for research purposes has been steadily increasing [2] with recent advances and diminishing technical barriers in data storage and information processing. There are, however, significant challenges in using EHRs effectively in the research setting and in ensuring the quality of data recorded for this purpose. Incorrect or missing data can render records as useless or indeed misleading such that conclusions drawn from the data could have a negative impact.

The aim of this chapter is to outline both the key challenges to the management and assessment of data quality in EHRs and the key considerations for meeting these challenges. The Clinical Practice Research Datalink database CPRD GOLD,

S. Dungey (✉)

Department of Informatics, University of Sussex, Brighton BN1 9QJ, UK

MHRA, 151 Buckingham Palace Road, London SW1W 9SZ, UK

e-mail: sheena.dungey@mhra.gsi.gov.uk

N. Beloff • A.R. Tate

Department of Informatics, University of Sussex, Brighton BN1 9QJ, UK

R. Williams • T. Williams • S. Puri

MHRA, 151 Buckingham Palace Road, London SW1W 9SZ, UK

globally recognised as being one of the largest and most detailed sources of electronic patient data, will be used as an example throughout. In Sect. 2, the concept of data quality is presented within the setting of primary care databases and a framework for its assessment is set out, based on findings of an investigation carried out on CPRD GOLD. In Sect. 3, the importance of understanding data quality of an individual source of data in relation to alternative sources, both intra- and internationally, is examined, posing the emerging challenges to the future use EHRs for research. Finally Sect. 4 investigates data quality requirements from the perspective of a range of stakeholders through discussion of a day-long CPRD-led data quality workshop and we consider the way forward to a more comprehensive approach to tackling issues of data quality in EHRs.

2 Developing a Data Quality Framework at the Clinical Practice Research Datalink (CPRD)

2.1 Defining Data Quality Within the Context of Electronic Healthcare Records

The widely accepted conceptualization of data quality is that it is defined through “fitness for use” [3] i.e. the ability of the data to meet the requirement of the user. To be able to use data to attain information, it must be complete, consistent and accurate and so forth. Actually defining what can be considered as complete or consistent, thus determining the quality of the data, will reflect the required use of the data and will therefore vary across purpose.

The primary purpose for recording patient data within the General Practice (GP) setting is to facilitate patient care and to assess and optimise the care of the practice population as well as to provide documentation for administrative and legal purposes. Guidelines are provided [4] to promote good recording practice including the use of codes to express clinical information and standard procedures for capturing information from outside the practice and with regards to sharing information. In 2004 the Quality and Outcomes Framework (QOF) was introduced with the provision of large financial incentives based on practice achievement on a range of quality of care indicators over 22 clinical areas [5]. This had a major impact on the use of clinical recording software and the development of a more standardised approach to data recording, particularly in coding of key disease areas [6] and recording of key lifestyle measures such as smoking status [7]. However, despite efforts to ensure quality and consistency in data recording, using GP records for research remains challenging [8].

The primary care setting is complex and constantly changing and this reflected in similarly complex and transient recording mechanisms. Coding systems, such as the Read code system predominantly used to categorise clinical events in UK primary care databases, including CPRD GOLD, are a prime example. Introduced to curb

the vast number of ways a clinical concept can be described, the number of codes has grown massively increasing risk of inconsistency in use and necessitating staff training [9, 10]. Further still, not all data is coded instead being entered as free text; such information is challenging to extract at the research stage, particularly amid growing concerns over record anonymisation [11, 12].

Additionally, some level of bias in data collection is hard to avoid. Whilst QOF has reaped many benefits to recording quality, these are tied to the clinical areas covered by QOF [13, 14]. Stigmatization of certain conditions is believed to lead to under-recording [15, 16] and the way data is recorded may depend on the type of staff entering the data and when it's entered relative to actual consultation [10]. And this is to name but a few examples.

Ultimately, the effort-benefit balance for detail of recording sits differently for patient care, for which GPs are striving to provide a face-to-face consultation in a time-pressurised environment, and for research, where meticulous and consistent recording is crucial.

Whilst the importance of addressing data quality in the reuse of EHR for research is widely acknowledged, with various frameworks having been put forward [17–19], there is no commonly recognised methodology for undertaking an assessment of data quality in this setting. Hitherto, data quality has been mainly addressed via one-off validation studies [20]. The evolution of a unified approach has no doubt been hampered by the vast number of possible measures and the variability in importance of measures between studies leading to data quality considerations derived for a given study being isolated to that study. A recent Clinical Practice Research Datalink (CPRD) sponsored project entitled “Methods to characterise and monitor data quality in the Clinical Practice Research Datalink” led by the University of Sussex, has addressed the need for standardisation and facilitation in data quality assessment. The rest of this section recounts the development of an approach for characterising data quality in primary care databases [6] based on investigative work carried out on the CPRD database.

2.2 Introduction to the Clinical Practice Research Datalink

The Clinical Practice Research Datalink (CPRD) GP OnLine Database (GOLD) contains diagnostic, demographic and prescribing information for over 14 million patients, broadly representative of the UK, providing a significant potential resource for public health and epidemiological research—its usage has led to over 1,500 published research studies and conference abstracts to date [21].

What is now known as CPRD was initially developed by an Essex general practitioner, Dr Alan Dean, to facilitate day-to-day management of his own general practice. This was so successful that a venture capital company was set up in 1987 named VAMP (Value Added Medical Products Ltd) to recruit other practices and form an information base. In late 1993, the company was taken over by Reuters and the database was offered to the Department of Health as independent custodian

to supervise access to the information for the benefit of public health. Early in 1994, the Office of Population Censuses and Statistics took over maintenance and running of the information resource (which was then renamed to General Practice Research Database), until 1999 at which point the Medicines Control Agency took over. Throughout this period, data collection and validation was maintained without loss of information from individual practices. This agency became the Medicines and Healthcare Products Regulatory Agency (MHRA) in 2003 following a merger with the Medical Devices Agency. Since then, use of the database expanded within the UK and overseas. In March 2011, the UK Government launched its “Plan for Growth” [22] which detailed steps needed to enable the British economy to become more internationally competitive. As part of this initiative the Government pledged to build a consensus on using e-health record data to create a unique position for the UK in health research. Under this motivation, CPRD was launched in April 2012 co-funded by the National Institute for Health Research (NIHR) and the MHRA.

CPRD is aiming at providing capability, products and services across a number of areas including secure integrated and linked data collection and provision, as well as advanced observational and innovative interventional research services [21, 23]. That the data is of high quality and can be validated as being so is paramount for all work carried out at CPRD.

2.3 CPRD GOLD Data Quality: Developing a Methodological Approach for Characterising Data Quality in Primary Care Research Databases

In order to ensure high quality data, CPRD historically has constructed a set of internal data quality measurements, at both patient and practice level. The practice level quality assessment is manifested by an ‘up-to-standard’ (UTS) date derived using a CPRD algorithm that looks primarily at practice death recording and gaps in the data. At patient level, records are labelled as ‘acceptable’ for use in research by a process that identifies and excludes patients with non-contiguous follow up or patients with poor data recording that raises suspicion as to the validity of that patient’s record. However, these checks are limited in scope, and with the expansion and increasing use of the database a more comprehensive approach is needed whereby CPRD undertake data quality assessment for these data both individually and jointly as linked data sets.

Work carried out under the current CPRD/University of Sussex data quality project commenced with a comprehensive examination of CPRD data quality and correlations between different measures, with a view to reducing the effective number of variables needed to characterise data quality [6]. This study was carried out as part of a wider project funded by the UK Technology Strategy Board and incorporated input from a user group consisting of representatives of pharmaceutical companies and clinical research organisations [23].

Several frameworks for data quality have been suggested in the literature. In our opinion, the use of different frameworks may not in itself be a major problem if clear definitions and examples are provided and all important aspects are considered. Ultimately, use of a framework is advantageous in encouraging a consistent and comprehensive approach to data quality assessment and hence, after a review of the literature, a suitable framework for describing dimensions of data quality was proposed. The dimensions include accuracy, validity, reliability, timeliness, relevance, completeness and integrity with full definition and examples given in [6]. Measures were then identified according to the framework and also as either basic or study-specific measures. Here, the definition as basic pertains to general measures such as recording of height and weight, duplicate records or missing values for fields such as staff ID. Definition as condition-specific pertains to measures characterising the coding of specific conditions. Note, all Clinical, Referral and Test event records have an associated Read code, as described above.

Table 1 shows the correlation matrix for selected basic measures. For most of the variables examined representing different aspects of patient records, correlations were very weak with (Spearman) correlation coefficients typically below 0.2 (absolute value). Most practices that were “bad” at recording one thing were almost always fine at recording all others. However, correlations between variables representing the same aspect were much higher. For example, percentages representing completeness of patient’s height, weight, smoking and alcohol status were found to be highly correlated (Pearson coefficient ≥ 0.79). The same was found to be true for study-specific measures for selected groups of patients, e.g. diabetes patients.

Additionally, the quality of coded data for research purposes (such as specificity and consistency of coding) in CPRD GOLD was found to be reasonably high for most of the criteria that we measured, especially in more recent years. Recording of most of the data elements that were investigated improved significantly between 2000 and 2010 with a noticeable improvement in 2004 for measures (such as those related to diabetes recording) that are included the Quality Outcomes Framework (QOF) introduced in that year.

2.4 Proposed Approach for Assessing Data Quality for Research

The fact that correlations between dissimilar variables are weak, representing the variability in recording for different criteria within each practice, leads to the necessity of an approach in which most of the data quality metrics are tailored to the intended use of the data. This approach was supported by the user group who agreed that some variables will be much more relevant to them than others, for example, the variables relating to the study-specific patient selection criteria. Additionally, study-specific variables are more likely to be intercorrelated and aggregation of variables into data quality summary scores becomes more feasible.

Table 1 Correlation matrix for general measures no time element—coefficients calculated using Stata 11 (StataCorp. 2009. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP) [9] extracted from 528 practices contributing data to CPRD over time period 2000–2010

Framework category	Variable	Valid reg. date	Valid age	Duplicates	Consultation date	Clinical event date	Referral date	Weight	Smoking	Alcohol
Validity	Valid reg. date	1.00								
Validity	Valid age	0.03	1.00							
Accuracy	Duplicates	0.01	0.03	1.00						
Validity	Cons. date	0.04	0.01	0.43	1.00					
Validity	Cli. event date	0.02	0.01	0.37	0.84	1.00				
Validity	Ref. date	-0.02	-0.09	-0.04	-0.19	-0.13	1.00			
Accuracy	Weight	-0.04	0.03	0.10	0.25	0.27	-0.03	1.00		
Accuracy	Smoking	-0.02	0.04	0.11	0.25	0.26	-0.02	0.83	1.00	
Accuracy	Alcohol	-0.05	0.08	0.11	0.21	0.24	-0.05	0.88	0.79	1.00

The disadvantage of this approach is that it may be necessary to measure data quality dynamically on a study-by-study basis (however, many criteria will be common across studies, e.g. completeness of recording of registration and life-style measures). To address this, computational methods for facilitating the dynamic calculation of study-specific measures are being investigated as part of the on-going work of the CPRD/University of Sussex data quality project.

It is proposed that basic checks are always carried out first for consistency of data elements between tables, duplicate values, missing values etc., before checking more complex elements. While this may seem obvious, in our experience these are often overlooked and even if the checks are carried out they are not often reported. It is also very important to investigate completeness and correctness of elements, such as dates and gender, as more complex elements will depend on these—for example if the registration dates of many patients are invalid then the incident rates will be flawed. Once these basic checks have been carried out, data quality measures based upon the intended use of the data can be derived via the following steps:

1. List all data elements required to define the cohort for the particular study, including all elements that these are dependent upon, e.g. registration and transfer out dates and specificity of coding of condition(s) of interest.
2. List all other elements that will be needed for the study, e.g. test results, smoking status, type of consultation.
3. According to the framework, determine data quality measures associated with each data element, specifying any conditions which must apply for a given data quality measure to be relevant. For example, in validating the coding pertaining to a condition, one can utilise the framework, working through the different components such as: **Accuracy**: are there coding errors? (e.g. a type 1 code for a patient diagnosed as type 2) and is there the coverage of expected associated tests? (e.g. HbA1c for diabetes). **Timeliness**: is the coding consistent over time? (e.g. consistent coding indicating the severity of a condition over time). **Relevance**: is the coding specific (e.g. the type of diabetes is given) . . . and so on.
4. It is proposed that the vast array of measures can be calculated using a contained set of core, input-driven computational routines. The underlying computations are identified as assessments of missing and implausible data, tested across the different natural structures of the database, namely within a given consultation record or over time. The inputs are a Read code list for a given condition or set of events and likewise for a set of associated events (e.g. diabetes and HbA1c test); the location of the entities of interest within the database (for HbA1c this would be the test table); a description of expected relationships between the data entities (here, as a gold standard, it is expected that three HbA1c should be recorded within 1 year). The proportion of patients failing each check can then be calculated at practice level.
5. Calculate incidence and prevalence rates for condition and check that these agree with data from the published literature and other sources. This step could be skipped if published validation studies exist.

6. Construct a set of indicators or scores for each practice. These could be the values of the practice based variables (i.e. the total number of fails of each measure for a given practice), or a combination of them (combination of data quality variables, as discussed above, is most likely possible for measures relating to a specific condition). The most appropriate method for combining variables into scores will depend on their intercorrelations and the intended use of the data. For more basic measures, simple thresholding could be applied for acceptable values. It is emphasised that scores should be used as a guide for further investigation rather than a hard and fast method of eliminating poor quality data. Exploring correlations and combining measures, thresholding measures and profiling patient sets (e.g. looking for underlying trends in poor quality such as patient age or particular ranges of test results) are essential components of an investigation into data quality.

2.5 Points for Further Consideration

In this study we investigated only coded data; however, additional information can be recorded in the free text. Free text is not widely used due to the cost and governance of anonymisation and wider difficulties of information extraction. However, free text could be highly valuable for validating coded information and for finding missing information that has not been coded [10, 24, 25], free text in relation to perceived data quality will be an important area of focus for future research. An interesting question being whether the use of free text could itself be used as an indicator of data quality relating to the completeness of recording.

Moving forward, it will be crucial to gauge the actual implications of poor quality to research outcomes to derive truly meaningful measures of quality. Whilst a measure of poor quality can indeed be defined as a shortcoming of the data in meeting a gold standard in data recording, this is not an absolute concept and does not necessarily predict consequences of including the poor quality data on the study outcomes derived from the data. This issue is currently being investigated by comparing outcomes from different data sets subject to a range of data quality constraints.

There is also a distinct need to understand the quality of the database compared with other sources of data. Although based on an investigation of a primary care database, much of our proposed approach would be equally applicable to other health care databases that are used for research, such as hospital records or registries, and also to linked data sets. The broader challenges involved in assessing multi-system data quality are the focus of the next section.

3 Emerging Challenges of Data Quality: Combing Data from Disparate Sources

3.1 Introduction

The potential advantages to research outcomes of inferring information from multiple sources, therefore extending the breadth and density of information available, are considerable [2]. However, if different sources of the same data are not comparable it indicates a quality issue intrinsic to one or both of the systems as a whole. If the limitations of data quality in different sources are not understood and addressed, poor quality may be amplified as a course of the linking process, manifesting as bias in conclusions drawn [26].

In this section, data quality issues relating to the combination of data from different national healthcare databases, from different domains of the UK healthcare system and from different data recording software systems are addressed, pursuing CPRD GOLD as an example, to pose what are widely perceived to be the key emerging challenges to the future use of EHRs for research.

3.2 *TRANSFoRM and the NIVEL Data Quality Framework*

The TRANSFoRm (Translational Research and Patient Safety in Europe) project [27] is an EU funded collaboration intended as a milestone project in the use of primary care EHRs for research. The aim is to provide interoperability between primary care databases (including CPRD GOLD) across Europe in order to facilitate research across resources, requiring common standards for data integration, data presentation, recording, scalability, and security. An extensive body of work has been carried out under the project; including a component to develop methodology for assessing and comparing primary care EHR data quality within different European databases, led by NIVEL (Netherlands Institute for Health Services Research) and in collaboration with CPRD [28].

Here we briefly present an example from the application of the NIVEL framework to the TRANSFoRm diabetes use case [29], the aim of which is to create a database of patients with type-2 diabetes containing genetic and phenotypic information compiled from genetic and primary care data sources (repositories). The combination of data from CPRD GOLD and NPCD¹ and derivation of a set of data quality measures were carried out according to the following steps. The study purpose and population as determined by the study purpose were defined for each database (via the individual national coding systems)—in this case the population of

¹The Netherlands national primary care database (NPCD), hosted by NIVEL, holds information from about 1.5 million patients (approximately 10 % of the total population).

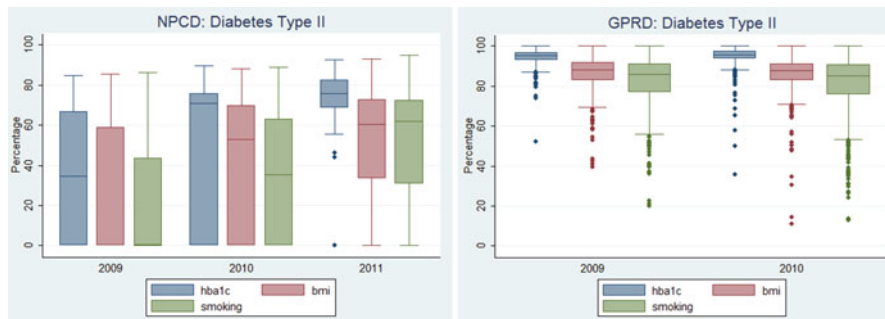


Fig. 1 Percentage of patients with type-2 diabetes at a given practice having at least one measurement of a given type within a given year for NPCR (300 practices) and CPRD (650 practices)

patients with type-2 diabetes. The data elements of interest to the researcher, such as related coded diagnoses and test results, were then set out. Finally, a quality assessment was made in terms of an evaluation of completeness, accuracy and correctness characteristics for each data element. An example is given in Fig. 1 of the completeness metric of three required data elements: measurements for hbA1c, smoking and weight [28]. Clearly, the degree of recording of these entities differs greatly between the two databases.

On comparison and integration of data derived from different national healthcare systems, data quality may be compromised by structural and operational differences in the healthcare systems. For one, there is variation across Europe in the extent that patient treatment is referred by the GP (largely the case in the UK). Whilst the Netherlands healthcare system is also based on the “Gatekeeper” model, patients in Holland are commonly seen by diabeticians rather than GPs which could account for the lack of measures for many patients. Differences in the data coding systems used present further challenges. The UK Read code system is greatly more complex than the ICPC coding system used across much of Europe, for example, for diabetes diagnosis there are only 2 standard ICPC codes compared to over 250 UK Read codes. Additionally the use of both coding systems has been shown to be variable within a given country and to be dependent on software package used to record data [30].

A wealth of other factors such as the age, set-up and management of a given database could be consequential in the recording of such events, as in Fig. 1, leading to patient information which is not, in the first instance, comparable across different databases. In the example of the diabetes use case, comparing prevalence between the two databases was problematic as the NIVEL database was much newer and thus some patients who had been diagnosed earlier may not be found. Key themes pertaining to data quality issues are summarised in Table 2 at the end of this section.

Ultimately this case study demonstrates the significant challenges faced on comparing just two European databases and the importance of ensuring a formalised, unified approach to assessing data quality for future linking of medical datasets across Europe.

Table 2 Factors affecting data quality, organised by stage of data usage from occurrence of the event to end-use of data

Data stage	Data stage factors	Potential effect on data quality
Occurrence of event	Organisational aspects of the health care system	For example, in a non-gatekeeping system (where patient treatment is referred through GP), an event may take place outside primary care, resulting in fewer event readings in primary care settings.
	Financial incentives in the health care system	The reimbursement system in one country may stimulate event readings—as occurs with the UK QOF system.
	Quality of care guidelines	For example, if a guideline says an event reading should be done every year, it will be more likely that such a measurement takes place.
Recording of event	Practice workload	Practice workload may have a negative effect on the carrying out of clinical events.
	Computerisation and EHR software	Studies [24] have shown there to be considerable differences between software packages in the way episodes of care are recorded.
	Strategic recording behaviour	Separate reimbursement schemes for patients with chronic illness will stimulate GPs to diagnose patients with chronic disease.
	Recording guidelines	Absence of recording guidelines will lead to less accurate, less complete and less correct data.
	The use of coding systems and free text	For example, the variety of ICDPC codes is much smaller than Read codes or Snomed coding system.
	Knowledge	Software packages and coding systems may enable GPs to record effectively, but if a GP is not trained, this may be of no use.
Practice workload	Practice workload	Shortage of time in a consultation will be detrimental recording behaviour.

(continued)

Table 2 (continued)

Data stage	Data stage factors	Potential effect on data quality
Data extraction	Extraction software Governance issues	Extraction software that determines which data elements will be extracted. Some patients/practice will opt out of data sharing.
Data import	Capacity of database to capture data extract Pre data entry quality control measures	Semantic interoperability may not be achievable across different software packages. This may for example mean that data that are incomplete are not entered into the database.
Generation of a research data file	Selection of data Linkage studies	Quality checks/filters may be employed such that not all data that is in a repository will go into the data file that is used by a researcher. Where data is linked, the resulting database will may hold only data on the common population.
Data analysis	Governance of repositories Choice of analysis method	There may be regulations restricting use of a certain repository for a certain purpose affecting completeness of data. Different methods of analysis as used by different researchers may render different results.

3.3 *Linked Data Sources*

Linking primary care data with data from different areas of UK healthcare is an evolving CPRD specialist service with data volume and coverage, as well as methodological expertise, seeing significant recent expansion. The need for a likewise development in classification and methodology for assessing data quality issues associated with linked data, and ensuring these considerations are a core component of the process of combining data from disparate sources, is now presented considering two CPRD linkage case studies.

3.3.1 *Cancer Recording and Mortality in the General Practice Research Database² and Linked Cancer Registries*

Boggon et al. [31] carried out an investigation into the completeness of case ascertainment in UK cancer registries collated into the National Cancer Data Repository (NCDR) by comparing information held within the NCDR to that of GPRD. UK cancer survival has been reported as being poor in international comparisons, however, it has been suggested that inaccuracies in cancer registration could invalidate international comparisons of cancer survival [32]. In particular, if patients with a good prognosis are missed or if patients are detected initially through death certification and if tracing backwards is inadequate, reported survival may be poorer than true survival.

At present, cancer registration is largely based on information supplied by hospitals and from death certification (via the Office for National Statistics). GPs in England do not routinely receive information directly from cancer registries and there has previously been little investigation comparing data from cancer registries and GP records [33, 34].

Boggon et al. found that on comparison of the two databases, levels of concordance between cancer registries and GPRD were reasonably high; however, numbers of patients known only to one dataset were non-trivial and levels of concordance in recording were observed to vary by cancer type. Overall survival rates were found to be higher in GPRD cancer cases. Ultimately the root of the recording disparities revealed in this study lies in the differences in how data is recorded for the two depositories such as when the data is recorded in relation to the event date and where the data is being generated in the first place.

²Study conducted before the 2012 transition from GPRD to CPRD GOLD.

3.3.2 Completeness and Diagnostic Validity of Recording Acute Myocardial Infarction Events in Primary Care, Hospital Care, Disease Registry, and National Mortality Records: Cohort Study

Herrett et al. [35] compared the incidence, recording, agreement of dates and codes, risk factors, and all-cause mortality of acute myocardial infarction recorded in four national health record sources encompassing primary care, hospital care, the national acute coronary syndrome registry, and the national death registry.

In over 2,000 patients, each data source missed a substantial proportion (25–50 %) of myocardial events. Again, missing data could be related to non-random features of the data life-span, particular to each data source. CPRD GOLD was the single most complete source of non-fatal myocardial infarction demonstrating the importance of incorporating primary care data in order to reduce biased estimates of incidence. Also highlighted here was the need for research into how electronic health record data are coded and how this can be improved. Additionally, it was concluded that more extensive cross referencing is required against additional sources of information on myocardial infarction, including investigation of electronic free text recorded by general practitioners (for example, diagnoses that are not recorded using a Read code).

Both the above examples demonstrate how data representing the same set of events but recorded in different settings can lead to different conclusions being drawn about those events. The value of linking these EHR databases is not that it will identify the gold standard or a superior database. Rather, the linkage will complement the information provided in each data source. An assessment of data quality is necessary to identify valid information within different sources and to understand why information may not be valid.

Conversely, the potential for comparison that is made possible through linkage enables greater insight into data quality issues. It may not be apparent that data is missing or incorrect until different sources of data are compared. Hence the increased utilisation of linkages will be important for development of more comprehensive and more fit-for-purpose data quality assessment procedures.

3.4 Comparing Data and Data Quality for Different GP Software Systems

Whilst CPRD GOLD data is currently collected from practices using a single software system (Vision), planned CPRD expansion will encompass data collection from additional data recording systems. In this scenario the reason for recording data is the same whichever source data is obtained from, unlike the examples discussed in Sect. 3.3, and is recorded under the same national system unlike the case discussed in Sect. 3.2. However, it would be naïve to dismiss the potential of the mechanism for recording of a given software system to affect recording

behaviour. UK systems are based around a problem-oriented model [36] i.e. events such as referrals and prescriptions are linked to a patient-problem, however, system-functionality can vary considerably under this premise [37]. Indeed [38] reported differences in the provision of primary care, as determined by levels of achievement on QOF indicators, dependent on the choice of software system. That variation is observed in the primary setting for data recording, without doubt warrants the investigation into the possible implications to the research setting.

To date there has not been a great deal of investigation into the impact of software system choice on research outcomes. Recently Reeves et al. [39] reported encouraging findings comparing the use of two primary care databases, CPRD (Vision software) and QResearch (EMIS software) [40], to carry out a study into the use of statins in patients with ischaemic heart disease. Whilst certain data characteristics were reported to vary between the different systems, the impact on the research findings was observed to be minor. However, as has been stressed throughout the course of this chapter, quality considerations vary greatly from study to study and more comprehensive examination is required. Understanding differences at this level may additionally aid a more fundamental understanding of the recording process at the point of care and help identify areas for focus in quality assessment.

3.5 Summary of Key Factors Affecting Data Quality

Key themes of the preceding sections are summarised in Table 2, based on the NIVEL/TRANSFoRm stepwise approach to data quality [28].

4 Report on 2013 Workshop on Data Quality

Throughout the preceding sections of this chapter, the scope of data quality in EHRs has been presented as wide-ranging, multi-faceted and dependent on use. Ultimately, the development of an approach for systematically assessing data quality for research needs to incorporate the expertise of the broad range of EHR users and contributors. In 2013, CPRD sponsored a one-day workshop entitled “Towards a common protocol for measuring and monitoring data quality in European primary care research databases” [41]. Researchers, clinicians and database experts, including representatives from primary care databases in the UK [6], Catalonia [42], Norway [43] and Spain [44] and from the Primary Care Information Services (PRIMIS) [45], were invited to give their perspectives on data quality and to exchange ideas on which data quality metrics should be made available to researchers.

The key findings from the discussions which took place throughout the day are summarised against the workshop objectives as follows:

1. Share experiences of assessing data quality in electronic health records (EHRs).

The expectation was that there would be a diverse set of perspectives from stakeholders. However, throughout the course of the day's discussions, there was a surprising amount of consensus, in particular concerning the characteristics that were important (particularly completeness, reliability and validity). Additionally presentations given on the day covered:

 - Approaches for assessing data quality for different national primary care databases and quality improvement through feedback mechanisms.
 - The impacts of poor data quality on research outcomes, the difficulty of revealing hidden quality issues and the time-consuming nature of data quality assessment and correction.
2. Discuss the issues and challenges involved with measuring data quality in EHRs for epidemiological and clinical research.
 - Clinicians highlighted that teams that generate personal health data are focused on clinical care rather than research.
 - Bias arises in areas that directly impact on reimbursement such as QOF [14]. Another potential influence is the role of the GP as patient advocate, where they might emphasise certain clinical findings to justify an investigation or referral.
 - Database managers highlighted the need for transparency of methods used to calculate variables and difficulties resulting when a data item can be recorded in multiple ways.
 - Those other than clinicians highlighted the need for a study-specific specific approach and that this makes a standard approach to assessment harder to achieve.
 - Data users stressed the importance of understanding unstructured/un-coded data and the ability to link data.
3. Work towards development of an approach to ensure compatibility of data quality measures for different European primary and secondary care databases.
 - Some technical proposals were put forward. A key discussion point for all groups was the communication of data quality metrics.
 - The importance of the publication of data quality work to inform the understanding of third parties, including future users of the data, regardless of whether this was the primary focus of a particular research study, or an early phase of data exploration. This should include how different parties handle the various aspects of data quality, the algorithms used for the identification of outcomes including code lists.
 - Most stakeholders agreed that data should be made available "warts and all" so users can make the decision on whether or not and how to use the data.
 - All agreed that it is important to have transparency on how the data is collected, and to understand the processes involved.

- Clinicians focussed on more specific examples such as test results and prescribing records.
 - Data experts discussed the benefits to quality that can be gained by being able to discuss data with clinicians.
4. Discuss how to help data contributors improve data quality (for both clinical care and research) at source.
- Providing practices feedback on their data quality, as pioneered by PRIMIS [45] in the UK and NOKLUS in Norway, has made an important contribution in motivating practices in high quality recording [46].
 - Data experts stressed the importance of providing recording guidelines to clinicians and feeding back data quality metrics directly to them.
 - It was generally agreed that incentivising GP's to produce higher quality data is key, either by feedback loops or by demonstrating how the data could be used to benefit their own patients. Clinicians are unlikely to prioritise data quality unless it benefits patient care or it affects their payments.
 - A critical success factor for the future will be to ensure that good quality data delivers value to those individuals who capture them, for example through decision support, alerts, charts of trends etc.
 - It may also be influential if clinical effort investments in data quality can be perceived as beneficial by patients themselves.
 - Further work is needed to understand the costs and benefits of improving data quality.

Based on the results of this workshop our suggestions for the moving forward are summarised below:

1. Data providers

- Provide meta-data and practice-based data quality scores to users (bearing in mind the concept of fitness for use).
- Be transparent about how data is handled providing as much information as possible on the processing steps.
- Provide information/training on how data is recorded at source.
- Explore ways to incentivise GPs to record better e.g. feedback data quality information or enabling database access for patient treatment.

2. Data users/experts

- Communicate impact of data quality on primary care data research to clinicians.
- Be aware of the limitations and impact of poor quality when carrying out research.
- Document or publish operational definitions so that researchers can easily validate research.

3. Clinicians

- Encourage training of staff within general practice to record data using coding as much as possible.

4. All

- Set up a network to continue the discussions of the workshop in order to develop a unified approach for measuring and improving data quality in Primary Care (and linked data) research databases.

Although the workshop did not result in a proposed overall approach for measuring data quality, many of the participants indicated that they would be interested in joining a data quality network to discuss these issues further. The network has recently been launched and will act as an international forum for discussion, aiming, through participation from different user groups, to develop a comprehensive, robust, integrated and widely used approach to measuring and delivering data quality across all aspects of EHRs (the reader is encouraged to contact the authors or further information about the network).

5 Conclusions

The issue of understanding data quality could not be more pertinent given the context of 'Big' healthcare data. Within England CPRD is the data service for healthcare research and has access to 10 % of the UK population at the primary care data level. It sits within a broader governmental project to make the entire health care data set available for use for the improvement of public health understanding and clinical research and delivery as well as surveillance. The principle component of this activity is a project from NHS England to gain access to primary care records from all General Practices in England, known as Care.Data [47]. Given this and increasing access to linkable data sets a key and full understanding of data quality within data sets from different platforms and collected for different purposes using different paradigms is central to use of this data for research and allied activities.

Furthermore, in the epidemiology and pharmacoepidemiology arena a growing number of projects are being developed to utilise big data from different global settings to answer key questions on drug safety for example. Projects such as OMOP (Observational Medical Outcomes Partnership) [48], OHDSI (Observational Health Data Sciences and Informatics) [49] and IMEDS (Innovation in Medical Evidence Development and Surveillance) [50] utilise disparate data sets in a common data model. The appropriateness of this approach needs to carefully consider the relative data quality aspects of each component data source.

Another focus of this chapter has been to convey the importance of collaboration in mapping out such an extensive and dynamic field, pursued here with the launch of a data quality network to facilitate the sharing of ideas. Understanding the requirements of and the demands upon clinicians must be central to addressing

quality issues in EHRs so as to be able to implement procedures that can realistically support data recording to facilitate patient management directly as well as to increase the capacity for research into cutting edge therapeutic and general patient care supporting longer term sustainability of the healthcare system.

References

1. Lawrenson, R., Williams, T., & Farmer, R. (1999). Clinical information for research; The use of general practice databases. *Journal of Public Health Medicine*, 21, 299–304.
2. Williams, T., van Staa, T., Puri, S., & Eaton, S. (2012). Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Therapeutic Advances in Drug Safety*, 3, 89–99.
3. Juran, J. M. (1988). *Juran's quality control handbook* (4th ed.). TX: McGraw-Hill.
4. The good practice guidelines for GP electronic patient records, v4. (2011). Provided by NHS/connecting for health.
5. National Institute for Health and Care Excellence. (2014). *Quality and outcomes framework*. Available via NICE <http://www.nice.org.uk/aboutnice/qof/indicators.jsp>. Accessed May 2014.
6. Tate, A. R., Beloff, N., Padmanabhan, S., Dungey, S., Williams, R., & Williams, T., et al. (2015). Developing a methodological approach for characterising data quality in primary care research databases (in press).
7. Taggar, J. S., Coleman, T., Lewis, S., & Szatkowski, L. (2012). The impact of the Quality and Outcomes Framework (QOF) on the recording of smoking targets in primary care medical records: Cross-sectional analyses from The Health Improvement Network (THIN) database. *BMC Public Health*, 12, 329–340.
8. de Lusignan, S., & van Weel, C. (2006). The use of routinely collected computer data for research in primary care: Opportunities and challenges. *Family Practice*, 23, 253–263.
9. de Lusignan, S. (2005). Codes, classifications, terminologies and nomenclatures: Definition, development and application in practice. *Informatics in Primary Care*, 13, 65–70.
10. Porcheret, M., Hughes, R., Evans, D., Jordan, K., Whitehurst, T., Ogden, H., et al. (2004). Data quality of general practice electronic health records: The impact of a program of assessments, feedback, and training. *Journal of the American Medical Informatics Association*, 11, 78–86.
11. Nicholson, A., Ford, E., Davies, K. A., Smith, H. E., Rait, G., Tate, A. R., et al. (2013). Optimising use of electronic health records to describe the presentation of rheumatoid arthritis in primary care: A strategy for developing code lists. *PLoS One*, 8(2), e54878.
12. Tate, A. R., Martin, A. G. R., Ali, A., & Cassell, J. A. (2011). Using free text information to explore how and when GPs code a diagnosis of ovarian cancer. Observational study using the General Practice Research database. *BMJ Open*, 1, 1–9.
13. Bhaskaran, K., Forbes, H. J., Douglas, I., Leon, D. A., & Smeeth, L. (2013). Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). *BMJ Open*, 3(9), e003669.
14. Mannion, R., & Braithwaite, J. (2012). Unintended consequences of performance measurement in healthcare: 20 salutary lessons from the English National Health Service. *Internal Medicine*, 42, 569–574.
15. Rait, G., Walters, K., Griffin, M., Buszewicz, M., Petersen, I., & Nazareth, I. (2009). Recent trends in the incidence of recorded depression in primary. *British Journal of Psychiatry*, 195, 520–524.
16. Salomon, R. M., Urbano Blackford, J., Rosenbloom, S. T., Seidel, S., Wright Clayton, E., Dilts, D. M., et al. (2010). Research paper: Openness of patients' reporting with use of electronic records: Psychiatric clinicians' views. *Journal of the American Medical Informatics Association*, 17, 54–60.

17. Gray Weiskopf, N., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20, 144–151.
18. Salati, M., Brunelli, A., Dahan, M., Rocco, G., Van Raemdonck, D. E., & Varela, G. (2011). Task-independent metrics to assess the data quality of medical registries using the European Society of Thoracic Surgeons (ESTS) Database. *European Journal of Cardio-Thoracic Surgery*, 40, 91–98.
19. Kahn, M. G., Raebel, M. A., Glanz, J. M., Riedlinger, K., & Steiner, J. F. (2012). A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical Care*, 50, S21–S29.
20. Herrett, E., Thomas, S. L., Schoonen, W. M., Smeeth, L., & Hall, A. J. (2010). Validation and validity of diagnoses in the General Practice Research Database: A systematic review. *British Journal of Clinical Pharmacology*, 69, 4–14.
21. Clinical Practice Research Datalink. <http://www.cprd.com>. Accessed May 2014.
22. HM Treasury. (2011). *Plan for growth*. Available via GOV.UK. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/221514/2011budget_growth.pdf. Accessed May 2014.
23. Tate, A. R., Beloff, N., Al-Radwan, B., Wickson, J., Puri, S., Williams, T., et al. (2014). Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface. *Journal of the American Medical Informatics Association*, 21, 292–298.
24. Woods, C. (2001). Impact of different definitions on estimates of accuracy of the diagnosis data in a clinical database. *Journal of Clinical Epidemiology*, 54, 782–788.
25. Wurst, K. E., Ephross, S. A., Loehr, J., Clark, D. W., & Guess, H. A. (2007). The utility of the general practice research database to examine selected congenital heart defects: A validation study. *Pharmacoepidemiology and Drug Safety*, 16, 867–877.
26. Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B., & Goldstein, H. (2014). Evaluating bias due to data linkage error in electronic healthcare records. *BMC Medical Research Methodology*, 14, 36.
27. TRANSFoRm. (2010). Website available at: <http://www.transformproject.eu>. Accessed May 2014.
28. Khan, N. A., McGilchrist, M., Padmanabhan, S., van Staa, T., & Verheij, R. A. (2013). *Deliverable 5.1: Data quality tool*. NIVEL, University of Dundee, CPRD. Available via TRANSFoRm. <http://transformproject.eu/Deliverables.html>. Accessed May 2014.
29. Leysen, P., Bastiaens, H., Van Royen, P., Agreus, L., & Andreasson, A. N. (2011). *Development of use cases*. University of Antwerp, Karolinska Institutet. Available via TRANSFoRm. http://transformproject.eu/D1.1Deliverable_List_files/DetailedUseCases_V2.1-2.pdf. Accessed May 2014.
30. de Lusignan, S., Minmagh, C., Kennedy, J., Zeimet, M., Bommeziijn, H., & Bryant, J. (2001). A survey to identify the clinical coding and classification systems currently in use across Europe. *Studies in Health Technology and Informatics*, 84, 86–89.
31. Boggon, R., Van Staa, T., Chapman, M., Gallagher, A. M., Hammad, T. A., & Richards, M. A. (2012). Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiology and Drug Safety*, 22, 168–175.
32. Beral, V., & Peto, R. (2010). UK cancer survival statistics are misleading and make survival worse than it is. *British Medical Journal*, 341, c4112.
33. Berkel, J. (1990). General practitioners and completeness of cancer registry. *Journal of Epidemiology and Community Health*, 44, 121–124.
34. Schouten, L. J., Höppener, P., van den Brandt, P. A., Knottnerus, J. A., & Jager, J. J. (1993). Completeness of cancer registration in Limburg, The Netherlands. *International Journal of Epidemiology*, 22, 369–376.

35. Herret, E., Dinesh Shah, A., Boggon, R., Denaxas, S., Smeeth, L., Van Staa, T., et al. (2013). Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: Cohort study. *British Medical Journal*, *346*, f2350.
36. Weed, L. (1968). Medical records that guide and teach. *New England Journal of Medicine*, *278*, 593–600.
37. Bossen, C. (2007). Evaluation of a computerized problem-oriented medical record in a hospital department: Does it support daily clinical practice? *International Journal of Medical Informatics*, *76*, 592–600.
38. Kontopantelis, E., Buchan, I., Reeves, D., Checkland, K., & Doran, T. (2013). Relationship between quality of care and choice of clinical computing system: Retrospective analysis of family practice performance under the UK's quality and outcomes framework. *BMJ Open*, *3*, e003190.
39. Reeves, D., Springate, D. A., Ashcroft, D. M., Ryan, R., Doran, T., Morris, R., et al. (2014). Can analyses of electronic patient records be independently and externally validated? The effect of statins on the mortality of patients with ischaemic heart disease: A cohort study with nested case-control analysis. *BMJ Open*, *4*, e004952.
40. QResearch. Website available at: <http://www.qresearch.org/>. Accessed May 2014.
41. Tate, A. R., Kalra, D., Boggon, R., Beloff, N., Puri, S., & Dungey, S., et al. (2014). *Data quality in European primary care research databases*. Report of a workshop sponsored by the CPRD in London September 2013. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2014, IEEE.
42. García-Gil Mdel, M., Hermosilla, E., Prieto-Alhambra, D., Fina, F., Rosell, M., Ramos, R., et al. (2011). Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). *Informatics in Primary Care*, *19*, 135–145.
43. Bellika, J. G., Hasvold, T., & Hartviysen, G. (2007). Propagation of program control: A tool for distributed disease surveillance. *International Journal of Medical Informatics*, *76*, 313–329.
44. Sáez, C., Martínez-Miranda, J., Robles, M., & Garcia-Gomez, J. M. (2012). Organizing data quality assessment of shifting biomedical data. *Studies in Health Technology and Informatics*, *180*, 721–725.
45. PRIMIS (Primary Care Information Services). Website available at: <http://www.nottingham.ac.uk/primis/index.aspx>. Accessed May 2014.
46. Lagerqvist, B., James, S. K., Stenestrand, U., Lindbäck, J., Nilsson, T., & Wallentin, L. (2007). Long-term outcomes with drug-eluting stents versus bare-metal stents in Sweden. *New England Journal of Medicine*, *356*, 1009–1019.
47. Care.Data. Website available at: <http://www.nhs.uk/nhsengland/thenhs/records/healthrecords/pages/care-data.aspx>. Accessed May 2014.
48. Observational Medical Outcomes Partnership. Website available at: <http://omop.org/>. Accessed May 2014.
49. Observational Health Data Sciences and Informatics. Website available at: <http://www.ohdsi.org/>. Accessed May 2014.
50. Innovation in Medical Evidence Development and Surveillance. Website available at: <http://imeds.reaganudall.org/>. Accessed May 2014.