A pragmatic approach for measuring data quality in primary care databases.

Sheena Dungey[1], Natalia Beloff[2], Shivani Puri[1], Rachael Boggon[1], Tim Williams[1] and A.Rosemary Tate[2]

***Abstract*— There is currently no widely recognised methodology for undertaking data quality assessment in electronic health records used for research. In an attempt to address this, we have developed a protocol for measuring and monitoring data quality in primary care research databases, whereby practice-based data quality measures are tailored to the intended use of the data. Our approach was informed by an in-depth investigation of aspects of data quality in the Clinical Practice Research Datalink Gold database and presentations of the results to data users. Although based on a primary care database, much of our proposed approach would be equally applicable to other health care databases.**

I. Introduction

Opportunities for using routinely collected health data for research purposes have increased enormously in recent years. In the UK, all primary care encounters are recorded electronically and practitioners are encouraged to make these records available for research. The Clinical Practice Research Datalink (CPRD) Gold database represents the largest collection of anonymised primary care patient records in the world. These data are used worldwide by academics, governments and the pharmaceutical industry for health services research. However, in common with most data that are collected for administrative or clinical purposes, data are variable in quality and may be missing or incomplete. We are developing a methodology for measuring data quality (DQ) in the CPRD Gold database which is focussed on the intended use of the data and on developing practice-based quality indicators which we will make available to both providers and users of the data.

II. Methods

In order to develop the protocol we first carried out an investigation of different aspects of DQ in the CPRD Gold database. We used the framework described below to ensure that all dimensions were covered. The selection of measures was carried out in consultation with our user group (which consisted of representatives of pharmaceutical companies and clinical research organisations). In this section we give details of the framework and briefly describe the methods for extracting data quality measures, which will be published elsewhere (manuscript in preparation).

*This work was supported by the Medicines and Healthcare Products Regulatory Agency

[1]Sheena Dungey, Shivani Puri, Rachael Boggon and Tim Williams are with the Medicines and Healthcare Products Regulatory Agency, Buckingham Palace Road, London, UK. Sheena.Dungey@mhra.gsi.gov.uk

[2]A.R.Tate and N. Beloff are with School of Informatics and Engineering, University of Sussex, Falmer BN1 9QJ, UK. rosemary@sussex.ac.uk

A. Framework

The literature on data quality is vast and, while there is general agreement regarding the definition and dimensions of data quality, there is much ambiguity in the terms that are used - not least in the medical field [1]. In March 2007, the UK Audit Commission published a framework to support improvement in data quality in the public sector [2]. This framework details six key characteristics (dimensions) of good quality data: Accuracy, Validity, Reliability, Timeliness, Relevance and Completeness. All can be applied to electronic health records. We modified the framework according to the hierarchical structure proposed by the authors of the Canadian Institute of Health framework [3] and added one more dimension; Integrity (Table I).

B. Data

All 629 practices that have been contributing data to the CPRD GOLD database using the Vision software on or after 01/01/1995 were considered for inclusion in the study. We excluded all patients who had only ever been temporarily registered with the practices and events that had been recorded to have occurred before the date of a patient's first permanent registration at the practice. For investigations over time we excluded practices that stopped contributing data prior to 01/01/2011 (n=91). Denominators for each year were based on all patients who were alive and registered at the practice for the full year in question, including patients with no events recorded in that year. From this dataset we extracted practice-based variables (percentages, means and medians) on different aspects of data quality for each practice. Examples of the variables include: percent of patients with a plausible weight record, (accuracy); percent of patients with valid registration date (validity); percent of diabetic patients with a code indicating diabetes type (relevance); non-missing data for test results (completeness). We investigated how these variables vary within and between practices and over time, using summary statistics and box plots to examine variation over time, and correlation analysis (Spearman ρ) and scatterplots to examine the relationships between pairs of variables.

III. Results

A. Results of investigation

538 of the 629 practices contributed data up until at least the end of 2010. The number of active patients who were registered for the whole year rose linearly from 3.9 million in 2000 to 4.2 million in 2010. The median number of patients per practice was approximately 6,700 (mean

TABLE I

PROPOSED FRAMEWORK FOR MEASURING DATA QUALITY IN PRIMARY CARE RESEARCH DATABASES (MODIFIED FROM THE UK AUDIT COMMISSION AND CIHI FRAMEWORKS)

Dimension /Characteristic	Criteria	Measure(s)
Accuracy		
Measurement or Recording Error	Implausible or incorrect values	N units with implausible/N units with value
Recording accuracy	Coding errors	N units with incorrect code /N units with code
	Recorded date is date event actually happened	% of units with incorrect date/ total units
Coverage	All events pertaining to a certain condition are recorded	N expected events/N actual events. Median number of relevant events recorded
	No duplicate records for same event	N duplicate units/total units
Validity		
	Correct/approved units or specified time intervals	N units with incorrect unit/N units, median time interval
	Valid proxy data	Validity score for codes obtained using data cleaning algorithms or from free text
Reliability		
Consistency/ Concordance	Information mismatch between various or within the same EHR data source	N units with mismatch/ N total units
Timeliness		
Practice level timeliness	Practice contributing data (to data centre) on time	Median time lag between date of record and date contributed
Timeliness of individual records	Practice recording events at time that they happened.	N events recorded late/ total events.
Relevance		
Coding specificity	to identify study population	N units with non-specific code/N units with code
Relevant time intervals	Time intervals adequate for the intended use	Median time interval for measure per unit
Completeness		
Unit non-response	Practice with no data for a period	Time period(s) with no data
Item non-response	Practice contributed no data for a certain element or only partial data	(N units for which data for a data element was not provided) /(N units that should have provided the data element)
Integrity	Control of provenance of data	Confirm data source and that data have not been changed or replaced in unauthorised way

7,200) in 2000 and 7,200 (Mean 7,700) in 2010. The quality of recording of most of the variables we measured was reasonably high (with median percentages for most variables of over 90%). However, there were large variations both between practices and over time, and most variables had left-skewed distributions with several outliers. Correlations between pairs of variables representing different aspects of data quality were very weak with very few having ρ above 0.2 (absolute value). 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 highly correlated with one another (Pearson r $\geq$ 0.79). The same was true for disease-specific measures for selected groups of patients, e.g. diabetes patients. Despite this wide variation, GP practices weak at recording one aspect were generally satisfactory at recording all others.

IV. SUGGESTED PROTOCOL

Initially, when we began this work, we hoped to be able to combine the variables representing various criteria into a smaller set of general indicators or scores which we could use to characterise each practice. However, the weak correlations between variables, (representing the variability in recording for different criteria within each practice) meant this was not feasible. We therefore decided that a better approach would be to tailor most of the data quality metrics to the intended use of the data. This decision was supported by our user group who pointed out that some variables will be much more relevant to them than others, for example variables relating to the study-specific patient selection criteria. An added advantage of this approach is that study-specific variables are more likely to be intercorrelated such that aggregation of variables into data quality summary scores becomes more feasible. The disadvantage is that it may be necessary to measure DQ dynamically on a study-by study basis. However, many criteria will be common to most studies. For example, completeness of recording of registration, birth and death dates will be important for cohort selection and calculation of incidence and prevalence rates. If the aim is to identify suitable practices at which to conduct clinical trials the validity of test results and completeness of recording of lifestyle measures will likewise be important.

A depiction of our suggested protocol is shown in Figure 1. We propose 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

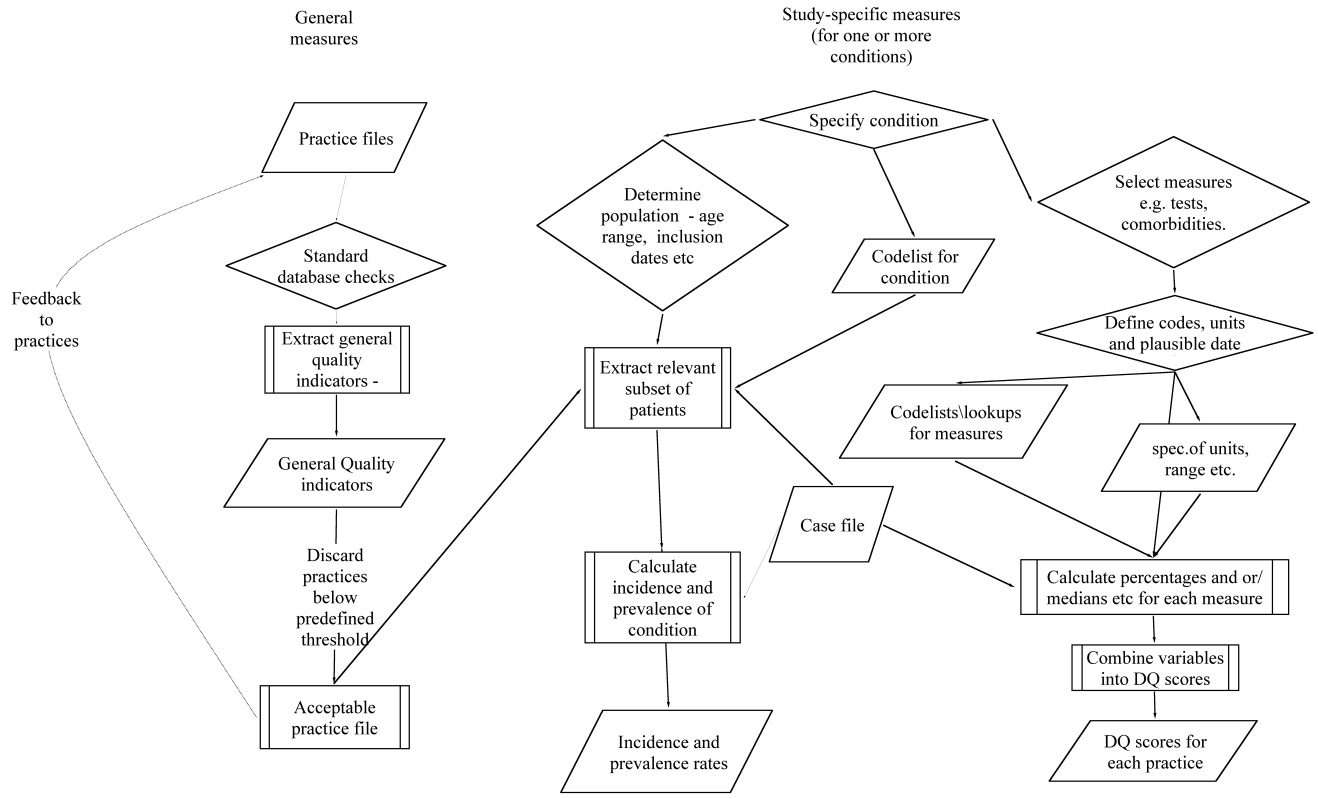

Fig. 1. Suggested protocol for extracting practice-based data quality measures for a specific study.

are invalid then the incident rates will be flawed. Once these basic checks have been carried out data quality measures can be based on the intended use of the data. We suggest the following two stage approach:

Basic general measures

Using the chosen Framework as a guide

1) Carry out basic database checks e.g. date validity, duplicates etc and make a note of practices below a certain threshold.
2) Carry out checks relating to timeliness for contributing data.
3) Check important aspects necessary for any study, e.g. gender (similar to the patient and patient checks currently conducted as standard by CPRD).

Study-specific measures

1) List all the data elements that are required to define the cohort for the particular study, including all elements that these are dependent upon, e.g. registration and transfer out dates etc. 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) Match the data quality modules to the list of data elements, specifying any conditions which must apply for a given data quality measure to be relevant. The proportion of patients failing each check can then be calculated at practice level.
4) Calculate incidence and prevalence rates for each 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.
5) Construct a set of indicators or scores for each practice. These could be the values of the practice based variables, or a combination of them. The most appropriate method for combining variables into scores will depend on their intercorrelations and the intended use of the data. For example, if the correlations are weak, weighted averages defined by the user could be constructed, whereas if they correlations are reasonably strong multivariate statistical methods such as principal components analysis should be considered. Another alternative, which we suggest for more basic measures would be simple thresholding on acceptable values.
6) Using the chosen Framework, check that all relevant dimensions have been covered as far as is practicably possible.

Although the study-specific variables may need to be calculated afresh for each new study, this can be implemented by writing generic programs which can take the various aspects of data quality as input.

V. Discussion

In this paper we present a pragmatic approach for measuring and monitoring data quality in large primary care databases. Development of the approach was informed by an in depth investigation of various aspects of data quality in

practices contributing data to a large primary care database, CPRD GOLD.

The lack of standardised methods for measuring data quality in medical databases is addressed in two recent studies which propose frameworks for evaluating data quality in medical database. Salati et al [4] use commonly used metrics from database research (accuracy, completeness, correctness, consistency and believability) [5], [6] to assess data quality in the European Society of Thoracic Surgeons (ESTS) Database. They suggest that, since this framework is data independent, it could be used as a template for measuring quality in other medical registries. More recently Kahn et al [7] emphasise the need for multisite data quality comparisons and propose "a more standardised and comprehensive approach" which incorporates and simplifies similar metrics (based on those suggested by [8] in a pragmatic approach).

Several other frameworks have been suggested for measuring data quality in clinical records (e.g. [3], [4], [7]). There is general agreement that quality depends on the intended use of the data, but no commonly agreed set of dimensions and criteria and indeed little agreement on the meaning of terms ([1]). 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 covered. However, if we wish to ensure a consistent approach for measuring data quality in clinical records we believe that it is important to have a consistent (and ideally commonly agreed) protocol.

We are currently implementing our approach in our large database of patient records and are using the proposed protocol to develop a suite of programs for extracting data quality statistics and constructing practice-based quality scores tailored to the intended use of the data. We shall use these to monitor data quality and to provide the results to research users. We plan to develop mechanisms to feedback the results to contributing practices in order to improve the data quality at source. We shall extend our current work to new sources of primary care data and then to the other datasets (secondary care, audit and clinical registry 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.

VI. Conclusions

Based on the findings of the investigations reported here it is concluded that understanding data quality as a multi-facetted and subjective concept which depends on the use that is being made of the data, i.e. fitness for use [9], is of central importance in deriving expedient data quality measures. It is also true, however, that a more uniform and consistent approach to data quality in healthcare databases is required. Different facets will be more important to some groups of users than others [10]. It therefore makes sense to incorporate information about the study into the derivation of data quality indicators which can be done by applying a set of generic data quality programs to a pre-defined list of data elements as specified by the user. We hope that publishing our protocol will stimulate research and discussion on developing a consistent approach to measuring data quality in research databases and that others might adopt and adapt the approach to their own databases.

References

[1] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research." *Journal Of The American Medical Informatics Association : Jamia*, vol. 20, no. 1, 2013.

[2] "Improving information to support decision making: standards for better quality data," Audit Commission, Report, 2007.

[3] CIHI, "Canadian Institute for Health Information, The CIHI Data Quality Framework, 2009 (Ottawa, Ont.: CIHI, 2009)," Canadian Institute for Health Information, Report, 2009.

[4] M. Salati, A. Brunelli, M. Dahan, G. Rocco, D. E. Van Raemdonck, G. Varela, and on behalf of the European Society of Thoracic Surgeons Database Committee, "Task-independent metrics to assess the data quality of medical registries using the European Society of Thoracic Surgeons (ESTS) Database," *Eur J Cardiothorac Surg*, vol. 40, no. 1, pp. 91–98, 2011.

[5] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *COMMUNICATIONS OF THE ACM*, vol. 45, no. 3, pp. 211–218, 2002.

[6] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for Data Quality Assessment and Improvement," *ACM COMPUTING SURVEYS*, vol. 41, no. 4, pp. 16:1–52, 2009.

[7] M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, and J. F. Steiner, "A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical research." *Medical Care*, vol. 50, Supplement, pp. S21–S29, 2012.

[8] R. Wang and D. Strong, "Beyond accuracy: what data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, March-May 1996.

[9] J. M. Juran, *Juran's Quality Control Handbook*, 4th ed. Mcgraw-Hill (Tx), 1988.

[10] C. Cappiello Cinzia, Francalanci and B. Pernici, "Data quality assessment from the user's perspective," in *Proceedings of the 2004 international workshop on Information quality in information systems*, ser. IQIS '04. New York, NY, USA: ACM, 2004, pp. 68–73.

Acknowledgment

This works was sponsored by the Clinical Practice Research Datalink and the UK Technology Strategy Board: project number 100926.