Data quality in European primary care research databases.

Report of a workshop held in London September 2013

A.Rosemary Tate[1] and Dipak Kalra[2] and Rachael Boggon[3] and Natalia Beloff[1] and Shivani Puri[3] and Tim Williams[3]

***Abstract*— Primary care research databases provide a significant resource for health services and epidemiological research. However since data are recorded primarily for clinical care their suitability for research may vary widely according to the research application or recording practices of individual general practitioners. A methodological approach for characterising data quality is required. We describe a one-day workshop entitled "Towards a common protocol for measuring and monitoring data quality in European primary care research databases". Researchers, database experts and clinicians were invited to give their perspectives on data quality and to exchange ideas on what data quality metrics should be made available to researchers. We report the main outcomes of this workshop, including a summary of the presentations and discussions and suggested way forward.**

I. Introduction

The potential for using routinely collected patient records for research purposes has been steadily increasing with the recent advances and diminishing technical barriers in data storage and information processing. Primary care records are created on, or close to, the date that an event occurs and record all interactions with the general practitioner (GP) including tests, prescriptions and referrals to secondary care. However, records are variable in quality and may be missing or incompletely recorded. Since the validity of results relies on the quality of the data, it is important to have processes in place for assessing this variability and ensuring that data is of high quality with respect to their intended use. Although there is a vast literature on data quality in general, and many different frameworks have been proposed, there is still a need to categorise different dimensions of quality and to standardise the benchmarks for each dimension.

Data quality is a multidimensional concept which depends on the use that is being made of the data, i.e. "fitness for use" [1]. Different dimensions will be more important for some groups of user than others. This workshop brought together clinicians, users of the data and database experts to discuss what data quality means to them and to develop a common approach for measuring data quality in primary care European databases. The specific aims were to:

1) Share experiences of assessing data quality in electronic health records (EHRs).
2) Discuss the issues and challenges involved with measuring data quality in EHRs for epidemiological and clinical research.
3) Work towards development of an approach to ensure compatibility of data quality measures for different European primary and secondary care databases.
4) Discuss how to help data contributors improve data quality (for both clinical care and research) at source.

The workshop was held at the Clinical Practice Research Datalink in London and was organised by the authors who chaired and facilitated the four sessions. These were arranged as two sets of short 10-minute presentations: A. Data quality in European research databases and B. Data quality from the users point of view and two discussion sessions: C. The clinical perspective (panel session) and D. Break-out discussions. The 42 invited attendees included statisticians, epidemiologists, general practitioners, clinician researchers, IT professionals and representatives from the Primary Care Information Services (PRIMIS). In advance of the workshop, all invitees were asked to provide answers to a questionnaire aimed at understanding what drives interest in data quality and how it is approached. We summarise the presentations, discussion and questionnaire answers and provide suggestions for a proposed way forward.

II. Summary of presentations

A. Data quality in European research databases

1) Data quality in the Clinical Practice Research Datalink (CPRD): Rosemary Tate described an investigation of data quality in the CPRD Gold database [2]. Percentages of data elements relating to different dimensions of data quality were extracted for all 538 practices contributing to the database between 2000-2011 and investigated using summary statistics, graphs and correlation analysis. Recording of most elements improved over time. There were large inter-practice variations, and most percentages had left-skewed distributions with several outliers. Most percentages were only weakly inter-correlated, except those related to specific conditions (e.g. tests and measures for diabetes). GP practices who were weak at recording one aspect were generally fine at recording all others. She concluded that practice-based DQ scores should be tailored to the intended use of the data.

2) Data quality in a primary care Catalan Database: Leonardo Méndez (SIDIAP) described the Registry Quality

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

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

[2]Dipak Kalra is with Centre for Health Informatics and Multiprofessional Education, University College London, UK. d.kalra@ucl.ac.uk

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

Score (RQS) scoring system that has been developed in order to select research-useable data from the SIDIAP database containing records from 274 primary care practices, representing 80% of the Catalan population [3]. The method is based on comparing disease prevalence rates in GP practices with the expected prevalence obtained either from literature or from "gold standards" such as disease or mortality registers. The prevalence rates were compared against those expected and used to assign scores to practices and staff. Only data above a certain score threshold was selected as research useable. These data were shown to be representative of the whole database and the Catalan population with respect to age, sex and geographical distribution; and have been validated for several diseases and mortality against gold standard registers. Recently introduced feedback to clinicians identifying patients with disagreement between new diagnosis, diagnosing criteria and prescriptions (poor data quality) produced a 66% improvement over 8 months. Feedback to clinicians is key to improve data quality.

3) Data quality in a Norwegian primary care database: Gustav Bellika (University of Tromso) described the state of the art in the use of primary care data for research in Norway and outlined the barriers that still need to be overcome. Although quality of recording in primary care is good for demographic, prescriptions and lab data, GPs use text, rather than codes. Gustav concluded that data quality can only be improved if providers use the data and that secondary users of the data have trust. They are deploying a surveillance system for infections diseases (SNOW) [4], which they hope will incentivise GPs to provide better quality data. Feedback loops to health professionals is believed to be essential for both improvement of coding practices, and clinical quality.

4) Data Quality Vector: Metrics for Biomedical Data Quality Assessment. Juan M Garcia-Gomez (IMBIE): Data quality studies have been based mainly on dimensions, procedures or requirements. However, there is a lack of consistency and generalization in current methods for biomedical data quality assessment, e.g. most studies focus on the semantic analysis and the use of health information standards and, in general, it is assumed that the probability distributions of data are static. The effects on data quality caused by data distributional shifts have been generally ignored. Juan presented a general framework, the Data Quality Vector (Saez et al, 2012), for the assessment of biomedical data quality, based in a set of probabilistic and semantic metrics associated with nine data quality dimensions (i.e predictive value, correctness, duplicity, consistency, completeness, contextualisation, temporal-stability, spatial-stability and reliability). This method is parametrizable and comparable across different domains, with special emphasis being put on the spatial and temporal stability of the data (Saez et al. 2013).

B. Data users and application of the data.

1) Data quality in epidemiological research: Liam Smeeth (LSHTM) discussed issues in data quality covering the different variables used in research, including outcomes, exposures and covariates. He provided examples showing how the implications of poor data quality and missing data can impact results. The impact will often be hidden, particularly if errors are differential across comparison groups, when bias can occur in any direction. Non-differential error will tend to bias towards the null leave residual (hidden) confounding. Specificity of outcome is of particular importance: false positive classification of outcomes will attenuate effect sizes and can obscure real beneficial or harmful effects.

2) Data quality and the primary care research network (PCRN): Greg Mickiewicz (PCRN). The PCRN provides infrastructure (e.g. research staff, funding to cover practice-based staff time, training) for conducting research in primary care settings and supports a wide range of research projects including: disease prevention, health promotion, screening and early diagnosis, as well as the management of long-term conditions. PCRN acts as the conduit for study teams to access patient-specific data, but does not hold such data. Greg described some of the implications of poor quality for the data which they do hold, which could include selecting the wrong sites for a clinical trial, failure to recruit to time and target, or over recruiting and making incorrect payments all with consequences for the study budget.

3) Data quality from the Pharmacoepidemiology point of view: Cathy Emmas (Astra Zeneca). Pharmaceutical companies routinely use primary care research database in their research. Cathy described some issues affecting data quality such as use of generic, non-specific or idiosyncratic Read codes or incomplete dosage information. She described the potential impact of poor data quality, such as extra time needed to assess DQ issues and to clean/reconstruct the data, the reduced sample size (and statistical power) due to incomplete data, and the validity of results, which could lead to missing disease and drug associations.

III. Panel discussion: Clinicians Perspective

Three clinicians (Dipak Kalra (UCL)), Tim Holt (University of Oxford) and Gro Berntsen (University of Tromso)) were asked for their position on three questions. Their answers and the ensuing discussions are summarised below

1) Given the many uses and users of primary care information, can there ever be a standard definition and metrics for data quality?: Quality is defined by ISO9000 as: "the degree to which a set of inherent characteristics fulfills requirement". In our experience different requirements for use of data may rely upon different characteristics of quality and in turn, different kinds of error may impact differently on different requirements. Since primary care information can be used for multiple purposes, there will never be a standard definition of its quality. The teams that generate personal health data are focused on clinical care rather than research, often with limited awareness of the data quality needs of quality assurance, governance, epidemiology, comparative effectiveness research etc. Although quality metrics can be defined, those of greatest importance to research, e.g. the accuracy of blood pressure recording and avoidance of rounding the observed values, might be difficult to impose

on clinicians unless they have an impact on the standard of patient care.

2) Do you see any 'conflicts of interest' between the different uses made of EHR information, that may impact on the accuracy and completeness of what is documented by clinicians?: There is an internationally recognised bias in clinical documentation in areas that directly impact on reimbursement, often called gaming [5]. English GPs might be influenced by payments associated with particular Quality Outcomes Framework (QOF) incentivised clinical codes, with un-measured areas of clinical work receiving less attention. 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.

3) What would be the persuasive influences (perhaps, incentives) that you would favour most to help improve the quality of clinical documentation?: Some aspects of data quality can be enforced by well designed clinical systems,e.g. by ensuring that numeric values are within physiological ranges, that clinical terms are appropriate to their context. However, system checks cannot normally detect incorrect but plausible values or be used to prevent fields being left blank (without annoying users). Providing practices feedback on their data quality, as pioneered by PRIMIS in the UK and NOKLUS in Norway, has made an important contribution to practices that are motivated to have good quality data: providing feed-back loops, which informs them in a non-punitive way about the quality of their work. This has positive effects on both data-quality and clinical quality [6].

However clinicians are unlikely to prioritise data quality unless it benefits patient care (at individual or practice population level) or it affects their payments. GPs take a pragmatic approach and coding behaviour (e.g. code specificity) is influenced by what is likely actually to influence management. 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 probably needed to understand the costs and benefits of improving data quality.

IV. Reports from breakout groups

Attendees were asked to classify themselves as either a clinician, data expert, data user or database manager and to form round table groups according to their classification. They were asked these two questions and given approximately 45 minutes to discuss.

1) What is the most important characteristic of data that determines its quality, from your perspective (for the type of data that you most often contribute or use)? : Three of the four groups (clinicians, data experts and data users) highlighted the importance of the breadth, or **completeness**, of the available data in order to interpret a specific piece of information in a broader context. The clinicians gave the examples of laboratory test results and prescribing records. The data users discussed the value of uncoded, unstructured text data, recorded by clinicians to support coded entries, and the potential to link patient records to other sources of data (e.g. CPRD GOLD primary care data linked to external secondary care, disease registry and mortality data). They also stressed the importance of being able to assess completeness on a study specific basis from an early stage in feasibility analysis and discussions. **Reliability**, in terms of data being consistent across an individuals patient record, was also raised by three groups (data experts, data users and database managers). Data users stressed the importance of being able to establish reliable start and end dates for patient follow-up in order to be able to follow patients over time. Data managers discussed the complexity introduced when the same data item can be recorded in a number of different ways, and the need to evaluate any consistent differences in the value of the original data item that may be dependent on the way it was entered. **Validity** was highlighted by three different groups from slightly different perspectives. The data experts discussed the value of being able to discuss the data with knowledgable clinicians who could establish the original meaning in order to avoid misinterpretation. The data users discussed the worth of being able to contact the data providers and conduct validation questionnaires around key events and records of interest. They also focused on the importance of external validity, in order to be able to generalise study results to broader population groups than those captured in the data itself. Other data characteristics mentioned by only one group included **accuracy** (clinicians), reasons for poor data quality (data users), **timeliness**, **integrity** over time, and **transparency** of methods used to calculate derived variables (database managers).

2) In your opinion, what would be the best approach to measure and report / represent data quality.: Data users and data managers agreed on the value of calculating and providing multi-dimensional aspects of data quality on a study specific basis, to build upon the current standard methods applied to individual patients and GP practices regardless of individual study requirements. It was also suggested that, due to variation in recording practices across clinicians within a GP practice, calculation of a data quality metric at the clinician level may further aid data users. Database managers suggested flagging incomplete records and implausible values, ideally as part of the data processing. Some of the broader concepts discussed in response to the first question were highlighted as key methods for measuring data quality, including research collaborations between data users and clinicians to aid interpretation of data items, triangulation of data sources through linkages methods. Validation studies, including incidence and prevalence calculated from the data source and compared to external sources, were also highlighted. Communication of data quality metrics and results more broadly was discussed by three groups (clinicians, data experts and data users). Data experts stressed the importance of providing recording guidelines to clinicians and feeding back data quality metrics directly to them, to support them in identifying areas for improvement. Others highlighted 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, and the algorithms used for the identification of outcomes, including codelists.

V. Summary of responses to the questionnaire

Completed questionnaires were received from 24 invitees which included some who were unable to attend. Participants were asked to specify their interest in data quality (classifications as in previous section) with the majority classifying themselves as data users. Responses to questions around key characteristics and the measuring and recording of data quality largely reflected the discussion in the breakout session as summarised above. The respondents generally recognised data quality as something fundamental to delivering to a high standard within their respective fields, be this to ensure reliable data for secondary use, to ensure high quality research or for the development of data standards for EHR. In terms of what participants hoped to gain from attending the workshop, there was a broad consensus of a need to gain insight into data quality issues in the wider data community. Interest was expressed in the different approaches and concerns both within groups of similar function and for groups providing a function on which ones work may rely, for example, for researchers understanding the data quality processes at the data collection stage.

VI. Conclusions

When this workshop was first proposed we had expected that different groups would have quite different perspectives on data quality. However, there was a surprising amount of consensus. 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. Although most researchers were mainly interested in aspects of data quality that were relevant to their own work, there was general consensus on the characteristics, that were important (particularly completeness, reliability and validity) and most agreed that data should be made available "warts and all" so users can make 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. Although there was awareness that primary care data has many limitations it was also generally agreed that they are extremely useful for research, provided that DQ issues are understood.

Although the workshop did not result in a proposed overall approach for measuring data quality, many of the participants (and questionnaire respondents) indicated that they would be interested in joining a European data quality network to discuss these issues further in order to develop a unified approach.

VII. Way forward

Based on the results of this workshop our suggestions for the way forward are summarised below.

1) Data providers
   - Provide meta-data and practice-based data quality scores to users, bearing in mind that DQ depends on the use of the data.
   - Be transparent about how data is handled and provide as much information on what has been done to the data - from start to finish.
   - Provide information/training on how data is recorded at source.
   - Explore ways to incentivise GPs to record better e.g. feedback data quality information. Explore ways to incentivise GPs by finding ways of letting them access the database to treat their patients.
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 European network which will continue the discussions of the workshop in order to develop a unified approach for measuring and improving data quality in European Primary Care (and linked data) research databases.

References

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

[2] T. Williams, T. VanStaa, S. Padmanabhan, and S. Eaton, "Recent advances in utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource," *Therapeutic Advances in Drug Safety*, 2012.

[3] M. D. Garcia-Gil, E. Hermosilla, D. Prieto-Alhambra *et al.*, "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*, vol. 19, no. 3, pp. 135–45, 2011.

[4] J. G. Bellika, T. Hasvold, and G. Hartviysen, "Propagation of program control: A tool for distributed disease surveillance," *INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS*, vol. 76, no. 4, pp. 313–329, APR 2007.

[5] R. Mannion and J. Braithwaite, "Unintended consequences of performance measurement in healthcare: 20 salutary lessons from the English National Health Service," *INTERNAL MEDICINE JOURNAL*, vol. 42, no. 5, pp. 569–574, MAY 2012.

[6] B. Lagerqvist, S. K. . James, U. Stenestrand *et al.*, "Long-term outcomes with drug-eluting stents versus bare-metal stents in Sweden," *New England J. Medicine*, vol. 356, no. 10, pp. 1009–1019, 2007.

Acknowledgment

This workshop was sponsored by the Clinical Practice Research Datalink. We would like to thank all the workshop speakers and participants for their valuable contributions and Sheena Dungey for summarising the questionnaire responses.