Profiling Clinical Datasets for Data Quality Assessment and Improvement

Wilfred Bonney, Donald Scobbie, Thomas Nind,
Sindy Donaldson-Buist, Christopher Hall, Emily Jefferson
Health Informatics Centre,
University of Dundee,
Scotland, U.K.
{w.bonney, d.scobbie, t.z.nind, s.x.donaldsonbuist, c.hall, e.r.jefferson}@dundee.ac.uk

Clinical datasets are the most critical resources or assets in the repository of Electronic Health Records (EHRs) and their quality gains competitive advantages in translational research. Accurate, reliable, and consistent representation of clinical datasets are essential for answering key research questions. However, a major issue with carrying out research on routinely collected primary care datasets is that they are often not fit-for-purpose or research-ready. It often takes months (if not years) for researchers to clean and transform clinical datasets for meaningful translational research. Profiling clinical datasets provides a proactive approach to examining and understanding the content, context and structure of source system data. The objective of this study was to develop a profiling dashboard to monitor, measure, assess, and improve the quality of clinical datasets hosted and maintained by the Health Informatics Centre (HIC) at the University of Dundee. Preliminary results indicated that the dashboard affords the flexibility to perform objective assessment of data quality, in terms of accessibility, accuracy, appropriate amount of data, completeness, and consistency.

Keywords: Clinical Datasets, Data Profiling, Data Quality, Translational Research.

1. INTRODUCTION

Routinely collected primary care datasets are certainly the most important resources for any population-based health outcomes. Clinical datasets not only serve as the resource for epidemiological studies, but they also enable researchers to appropriately answer research questions (Abhyankar et al., 2012; Bonney et al. 2014). Health outcomes based on the secondary uses of routinely collected primary care datasets will ultimately be flawed if the underlying data quality is poor and/or error-prone. Accurate, reliable, and consistent representation of clinical datasets are essential for answering key research questions. However, a major issue with carrying out research on routinely collected primary care datasets is that they are often not fit-for-purpose or research-ready.

Clinical and biomedical researchers continue to struggle in cleaning and transforming clinical datasets before they are made fit-for-purpose. In order for primary care datasets to be meaningful and useful for translational research and other secondary uses, the quality of the datasets must be assessed. The pervasive nature of primary care datasets makes them good candidates for applying data profiling techniques to assess their quality (Batini et al., 2009). Data profiling provides the platform to assess and monitor the quality of clinical datasets.

Data profiling, also known as *data discovery* or *data auditing*, is specifically about discovering data and the characteristics of that data (DataFlux Corporation, 2003). It provides a proactive approach to understanding the content, context and structure of data (Comingore, 2008; DataFlux Corporation, 2003; Eckerson, 2004; Mansingh and Srikant, 2005). Comingore (2008) explicitly defined data profiling as the "process of examining source system data and collecting various statistics for data quality, data integration and data augmentation assessment" (p. 21). The collection of various statistics from data profiling makes it a useful toolkit for assessing the quality of clinical datasets. Studying and analysing the collected statistical data also provides the opportunity to validate them against any expected data formats and values (Miriyala, 2007). More importantly, the statistical data gathered during the data profiling process could be aggregated and used to assess and improve the quality of clinical datasets.

The quality of information needed for evidence-based medicine and decision support is heavily influenced by the underlying data quality (Shariat Panahy et al., 2013).

According to Wang and Strong (1996), improving data quality requires understanding what data quality means to data consumers. However, there is a lack of a unified definition for data quality. Citing the work of Wang (1998), Liaw et al. (2012) defined data quality in the context of fit-for-purpose or fit-for-use. Karr, Sanil and Banks (2006), on the other hand, defined data quality as the "capability of data to be used effectively, economically and rapidly to inform and evaluate decisions" (p. 138).

Karr et al. (2006) embodied data quality as a decision problem by asserting that the quality of data decreases with increasing quantity of data. This decision problem intensifies and becomes crucial when the "quality of data maintained by healthcare organizations is becoming a critical factor in the delivery of medical care" (Lorence, 2003, p. 425). Hence, the need for high quality datasets has never been greater in the healthcare industry (Mphatswe et al., 2012).

Clinical datasets are only fit for secondary uses if they are free of defects and possess desired quality features (Redman, 2001; Shariat Panahy et al., 2013; Weiskopf & Weng 2013). Redman (2001) unequivocally asserted that "data are of high quality if they are fit for their intended uses in operations, decision-making, and planning" (p. 241). In other words, clinical datasets are only as useful as their desired quality (Choquet et al., 2013; Lorence 2003; Orfanidis et al., 2004; Svensson-Ranallo et al. 2011). Therefore, the need to ensure that clinical datasets, extracted for research purposes, are of high quality is of great essence.

The Health Informatics Centre (HIC) at the University of Dundee not only provides a service to securely host clinical datasets, but it also provides pseudonymised extracts of cohorts to researchers so as to enable them answer key research questions (with appropriate governance approvals).

Historically, each research group receiving a dataset will investigate the data provenance to identify gaps and/or errors within the dataset. This exercise takes considerable time and if it is not done systematically, errors and inconsistencies will be introduced and/or missed. If different research groups analyse the data with a different understanding of the data cleanness or data errors, different conclusions can be inferred from the data. The objective of this paper was to develop a profiling dashboard to monitor, measure, assess, and improve the quality of clinical datasets hosted and maintained by HIC.

2. MATERIALS AND METHODS

The approach involved the utilisation of three major data profiling techniques, recommended by DataFlux Corporation (2003), to develop a data quality dashboard to monitor and report on the quality of HIC-hosted clinical datasets. The three techniques consisted of: (a) structure discovery (i.e. understanding data patterns and metadata); (b) data discovery (i.e. discovery of business rule validations, and data accuracy and completeness); and (c) relationship discovery (i.e. discovery of data similarity and redundancy) (DataFlux Corporation 2003). The three techniques were applied on three datasets (i.e. Biochemistry, Prescribing and SMR01) hosted and maintained by HIC. The three datasets are described below:

- Biochemistry Dataset: This dataset contains all the Fife and Tayside biochemistry tests.
- Prescribing Dataset: This dataset contains all the community-dispensed prescription data from Fife and Tayside.
- SMR01: This is a Scottish Morbidity Record (SMR) dataset that collects episode level data on hospital inpatient and day case charges from acute specialities from hospitals in Fife and Tayside, Scotland.

The methods also involved the identification and quantification of data quality dimensions/metrics appropriate for the datasets. The data quality metrics identified for this study were derived from existing data quality frameworks developed by Almutiry, Wills, and Crowder (2013); CIHI (2009); Liaw et al. (2012); and Wang and Strong (1996). The harmonisation of the frameworks was necessary to ensure that the identified data quality dimensions or metrics were in alignment with the process of managing clinical datasets stored in the repository of EHRs. Through the harmonisation process, some key data quality dimensions such as timeliness, currency and volatility were deemed to be not appropriate for the HIC-managed datasets. For example, timeliness (i.e. the degree to which the "age of the data is appropriate for the task at hand" (Wang & Strong, p. 32)) was not considered as critical factor in the assessment process as the datasets in question are longitudinal in nature.

Review of relevant literature revealed several approaches to assessing data quality (Batini et al., 2009; CIHI, 2009; Liaw et al., 2012; Naumann and Rolker, 2000; Pipino et al., 2002). Acknowledging the fact that data quality is a multi-dimensional concept, Pipino et al. (2002) recommended that, in assessing data quality, organisations "must deal with both the subjective perceptions of the individuals involved with the data, and the objective

measurements based on the data set in question" (p. 211). In other words, assessing data quality requires both subjective and objective assessments. Whereas the subjective assessments reflect the needs and experiences of stakeholders, the objective assessments reflect states of the data with or without the contextual knowledge of the residing application (Pipino et al., 2002).

For the purposes of this study, objective assessments were used to quantify and measure the identified data quality metrics shown in Table 1. Table 1 depicts the operationalisation of the data quality metrics used in the study, categorised into the four data quality framework proposed by Wang and Strong (1996).

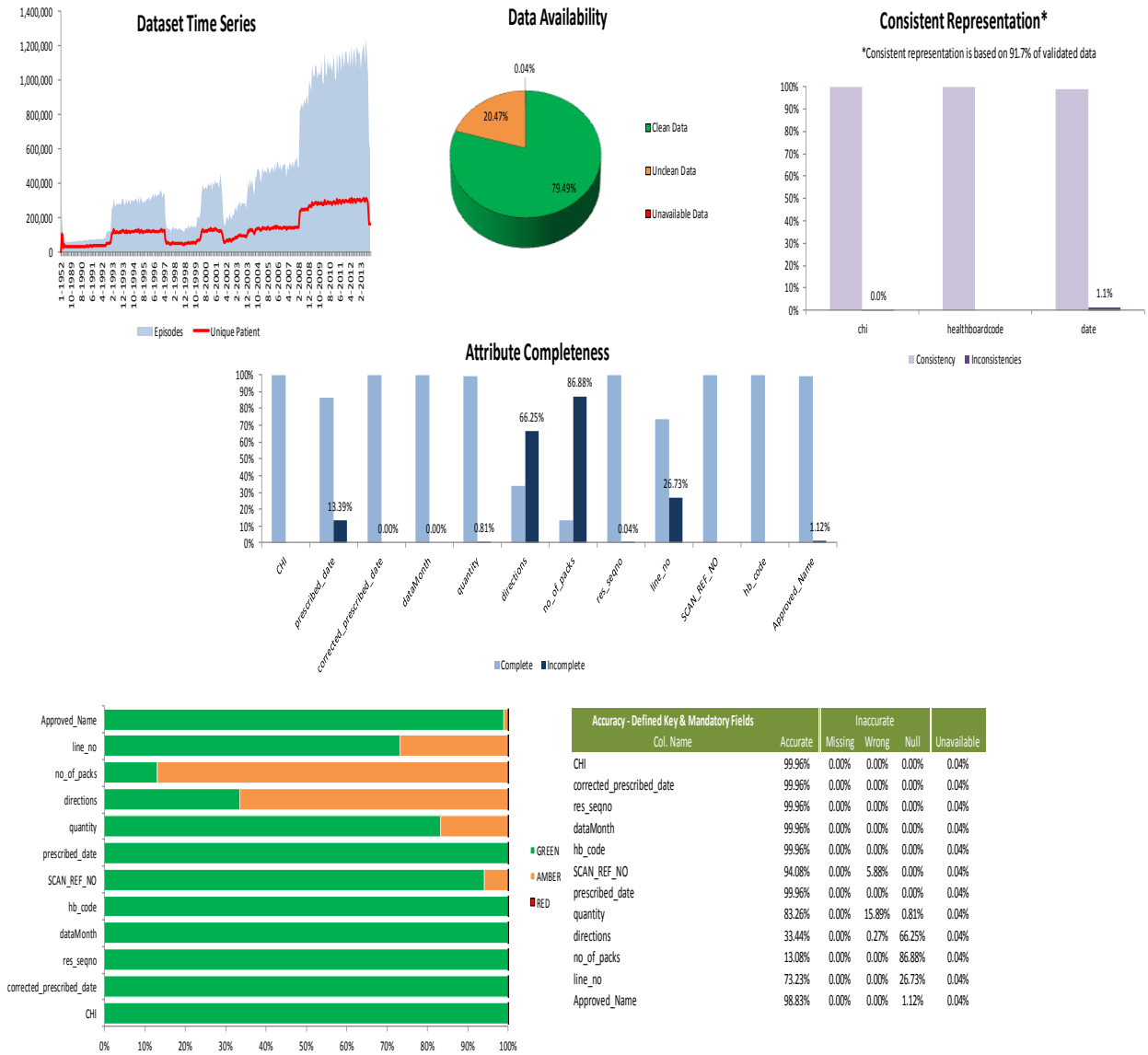
3. RESULTS

The profiling dashboard provided a platform to monitor and report on the quality of HIC-managed datasets. Preliminary results indicated that the dashboard affords the flexibility to perform objective assessment of data quality, in terms of accessibility, accuracy, appropriate amount of data, completeness, and consistency (Batini et al. 2009; Liaw et al. 2012; Pipino et al. 2002; Strong et al. 1997; Wang and Strong 1996). Figure 1 depicts the current view of the data quality dashboard for the Prescribing dataset.

As can be viewed from the dashboard in Figure 1, it provides the unique information about the dataset, depicting the time series and data availability in the context of clean data, unclean data and unavailable data. In its current form, 79.49% of the Prescribing dataset is clean and 20.47% of the data is unclean and requires further cleaning. There is also 0.04% of the data that is not made available to researchers. This unavailability is due to issues with the data that cannot be easily resolved as a result of legacy systems. The essence of accuracy, completeness and consistent representation is discussed in section 4 of this paper.

Table 1: *Identified data quality metrics for the study*

Data Quality (DQ) Category	**DQ Dimension/Metric**	**Definition (Wang and Strong, 1996)**	**Measure (Batini et al., 2009, p. 19)**
Intrinsic DQ	Accuracy	The degree to which "data are correct, reliable, and certified free of error" (p. 31)	Syntactic Accuracy = Number of correct values/number of total values
Accessibility DQ	Accessibility	The degree to which "data are available or easily and quickly retrievable" (p. 32).	Accessibility = max (0; 1-(Delivery time - Request time)/(Deadline time - Request time))
Contextual DQ	Appropriate amount of Data	The degree to which "the quantity or volume of available data is appropriate" (p.32).	Appropriate Amount of data = Min ((Number of data units provided/Number of data units needed); (Number of data units needed/Number of data units provided))
	Completeness	The degree to which data are not missing and are of "sufficient breadth, depth, and scope for the task at hand" (p. 32).	Completeness = Number of not null values/total number of values
Representational DQ	Consistent Representation (Consistency)	The degree to which "data are always presented in the same format" (p. 32).	Consistency = Number of consistent values/number of total values



Accuracy - Defined Key & Mandatory Fields		Inaccurate			
Col. Name	Accurate	Missing	Wrong	Null	Unavailable
CHI	99.96%	0.00%	0.00%	0.00%	0.04%
corrected_prescribed_date	99.96%	0.00%	0.00%	0.00%	0.04%
res_seqno	99.96%	0.00%	0.00%	0.00%	0.04%
dataMonth	99.96%	0.00%	0.00%	0.00%	0.04%
hb_code	99.96%	0.00%	0.00%	0.00%	0.04%
SCAN_REF_NO	94.08%	0.00%	5.88%	0.00%	0.04%
prescribed_date	99.96%	0.00%	0.00%	0.00%	0.04%
quantity	83.26%	0.00%	15.89%	0.81%	0.04%
directions	33.44%	0.00%	0.27%	66.25%	0.04%
no_of_packs	13.08%	0.00%	0.00%	86.88%	0.04%
line_no	73.23%	0.00%	0.00%	26.73%	0.04%
Approved_Name	98.83%	0.00%	0.00%	1.12%	0.04%

Figure 1: *Profiling dashboard for Prescribing dataset*

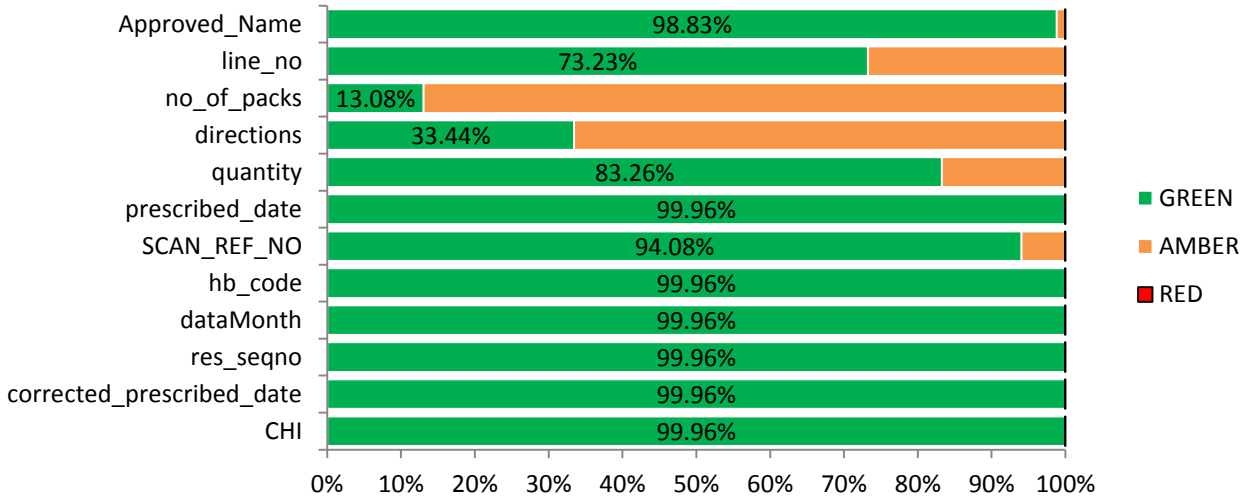


Figure 2: *Syntactic accuracy of Prescribing dataset*

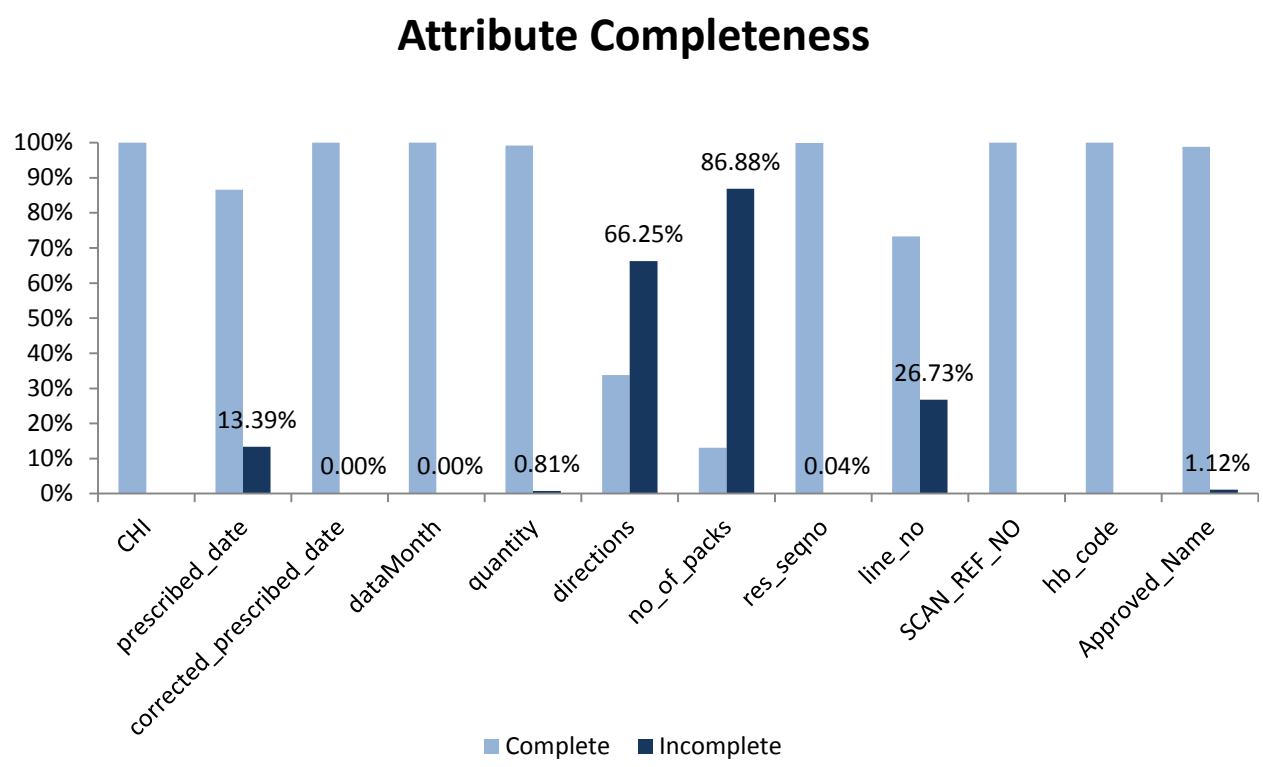


Figure 3: *Attribute Completeness of Prescribing dataset*

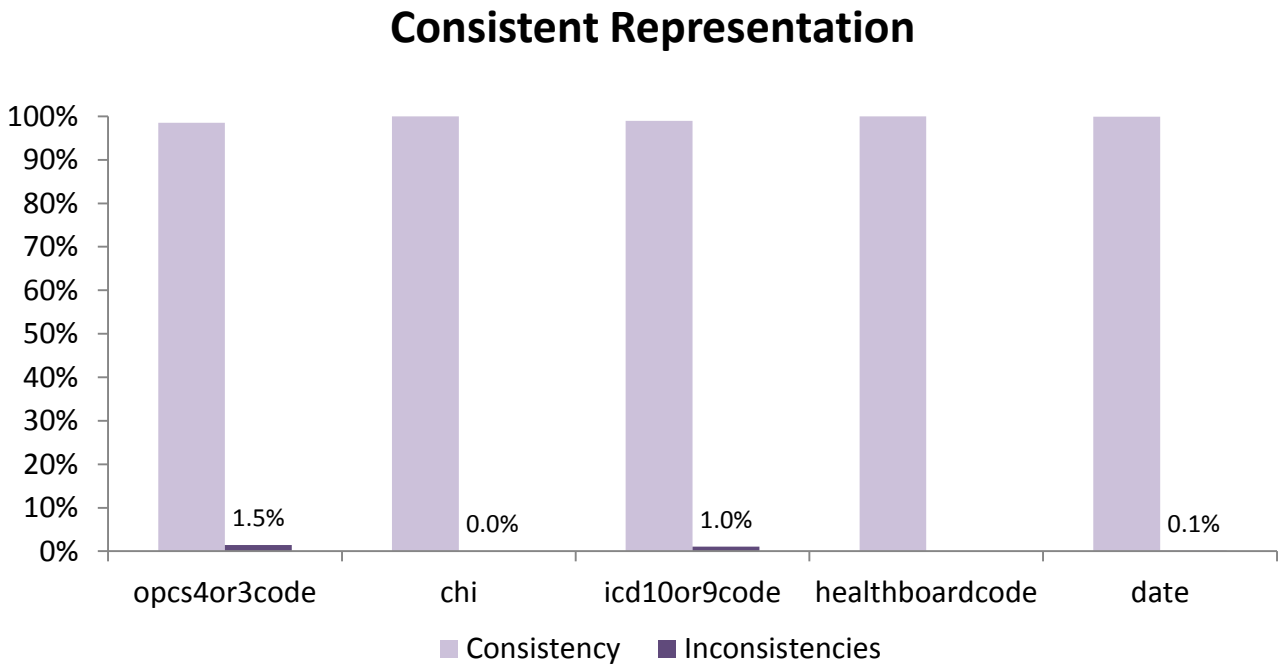


Figure 4: *Consistent representation of SMR01 dataset*

4. DISCUSSION

This section discusses the essence of assessing *Accuracy*, *Completeness* and *Consistent Representation* in the HIC-managed datasets.

4.1 Accuracy

The accuracy measure (show in Table 1) was used to measure the syntactic accuracy of the HIC-managed datasets. Figure 2 shows the detailed accuracies for the attributes in the Prescribing dataset. It is obvious from the output that the accuracies of most of the attributes are in excellent condition. The colour coded RED, AMBER and GREEN has the following meaning:

- RED denotes the percentage of unavailable data. In this case, 0.04% of the data is unavailable.
- AMBER denotes the percentage of data that is missing, wrong, and/or null based on HIC validation rules.
- GREEN denotes the percentage of data defined as accurate based on HIC validation rules.

It is also clear from Figure 2 that the attributes: *no_of_packs* and *directions* have accuracy issues within the Prescribing dataset, as their accuracies are respectively 13.08% and 33.44%. These low values in accuracy are attributed to the fact that the *no_of_packs* field data was historically populated between 1997 and 2000. However, it was determined after 2000 that the field was not useful for research and there was no incentive to populate it. The population of the field therefore ceased after 2000 and has since contributed to a high number of nulls in the *no_of_packs* field and thus accounts for the low accuracy measure within the Prescribing dataset. Similarly, in the case of *directions,* it was determined in 2002 that it was not cost-effective to manually populate the field in the dataset, as the field was only used in a small number of studies. Specifically, there was a trade-off between the cost of entering the values in the dataset and the usage of the field by researchers. HIC continues to populate the field as per project request from researchers. This might account for the slight increase in accuracy in the *directions* field compared to the *no_of_packs* field.

4.2 Completeness

As in the case of accuracy, the attribute completeness for the Prescribing dataset (shown in Figure 3) also indicates that most of the *no_of_packs* and *directions* fields' data are sparsely populated. The attribute completeness for *no_of_packs* and *directions* were respectively 13.12% and 33.75%.

The reason for these low values in completeness is as explained in section 4.1.

The use of these two attributes in research will undoubtedly skew the results to the left or right, depending on the key research question. The output from Figure 4 is therefore very useful in informing HIC Data Analysts (i.e. those within HIC who generate the data extracts used for research) that careful consideration should be taken when releasing datasets to researchers, especially, in the case when attributes' data have very low completeness values. The onus is on the HIC Data Analysts to ensure that researchers are warned in advance about any low completeness values. The researchers can then make an informed decision whether to use or ignore those fields in their research.

4.3 Consistent Representation

The consistent representation result (shown in Figure 4) relates to both internal consistency with Date, Community Health Index (CHI), Health Board Codes; and external consistency with International Classification of Diseases (ICD-9/10), and Office of Population Censuses and Surveys (OPCS) *Classification of Surgical Operations* OPCS-3/4 within the SMR01 dataset. Overall, the dataset consistency with ICD-9/10 was much higher than that of the OPCS-3/4. Whereas 1% of the ICD-9/10 codes require further cleaning, 1.5% of the OPCS-3/4 codes needed further cleaning in order to improve the quality of the SMR01 dataset.

The inconsistent representations of ICD-9/10 and the OPCS-3/4 codes within the SMR01 dataset definitely have implications when it comes to using the dataset for research. Researchers relying solely on these codes for research might need to be aware of the proportion of the inconsistencies and can then make an informed decision to ignore those records with inconsistent codes in their copy of the dataset. Improving the data quality of the SMR01 dataset will require that the inconsistencies are resolved before making the dataset available to researchers. Alternatively, the dataset could be released to researchers without including the records with inconsistent representation of ICD-9/10 and the OPCS-3/4 codes.

5. CONCLUSION

Data quality remains a major concern in the healthcare industry when it comes to the secondary uses of primary care datasets. The health of clinical datasets depends enormously on their quality. Health outcomes based on clinical datasets will ultimately be flawed if the underlying data quality of

the datasets is poor and ineffective. Hence, every effort should be made to ensure that clinical datasets, extracted for research purposes, are of a high quality with minimum or no error-prone.

Profiling and assessing the quality of clinical datasets are important because their quality gains competitive advantages in translational research. The study has provided a framework for profiling and assessing the quality of clinical datasets. The developed profiling dashboard is scalable and reproducible, and provides a consistent and systematic approach to assessing data quality across all datasets and attributes. The tool is currently utilised by HIC Data Analysts to internally monitor on the quality of the HIC-managed datasets and to identify opportunities for data cleaning and quality improvements. It promptly warns HIC Data Analysts in cases where newly loaded data is presenting statistically significant deviations from the norm. The tool is also used to share with research groups on the provenance of the HIC-hosted clinical datasets.

Developments are underway so that the profiling dashboard can also be utilised on top of each research extract. The cleanliness of the data will be different for each specific research extract depending on the cohort and attributes required by the project. Using this tool, the work to measure accuracy, completeness and consistency across all attributes will not need to be carried out by each research group.

The profiling dashboard could then be presented to researchers scoped within the requested data for them to quickly analyse what to expect and to minimise the time usually spent running the analysis themselves. There is also continuing plan to integrate the generated profiling reports with DataCleaner Monitor (DataCleaner 2014) and/or QlikView (QlikView, 2014) to facilitate interactive and real-time monitoring and reporting on the quality of the HIC-managed datasets. Future work will also involve performing the subjective assessments of data quality. This will provide a platform to monitor, measure, and assess the data quality of HIC-managed datasets, in terms of usability, interpretability and relevancy. The subjective assessments will ensure that feedback from the HIC data users group is used as a tool to improve the quality of the datasets. More importantly, the combination of the objective and subjective assessments will ultimately improve the quality of clinical datasets released to researchers by HIC.

6. ACKNOWLEDGMENTS

The authors acknowledge the support of the Health Informatics Centre (HIC), University of Dundee for managing and supplying the datasets and NHS Tayside, the original data source.

7. REFERENCES

Abhyankar, S., Demner-Fushman, D. and McDonald, C. J. (2012) Standardizing clinical laboratory data for secondary use. *J. Biomed. Inform.,* 45(4). 642-650.

Almutiry, O., Wills, G. and Crowder, R. (2013) Towards a framework for data quality in electronic health records. In *Proceedings of IADIS International Conference, e-Society 2013, Lisbon, Portugal.*

Batini, C., Cappiello. C., Francalanci, C. and Maurino, A. (2009) Methodologies for data quality assessment and improvement. *ACM Comput. Surv*., 41(3). 1-52.

Bonney, W., Doney, A. and Jefferson, E. (2014) Standardizing biochemistry dataset for medical research. In: Bienkiewicz M, Verdier C, Plantier G, Schultz T, Fred A, Gamboa H, Editors. *Proceedings of HEALTHINF 2014: International Conference on Health Informatics*. Angers, France, 3-6 March 2014. Portugal: SciTePress. 205-210.

Canadian Institute for Health Information (CIHI). (2009) *The CIHI Data Quality Framework*. Available from http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en (12 March 2014).

Comingore, D. (2008) SQL server 2008 data profiling. *SQL Server Mag.*, 10(7). 21-25.

Choquet, R., Qouiyd, S., Ouagne, D., Pasche, E., Daniel, C., Boussaïd, O., and Jaulent, M. (2010) The information quality triangle: A methodology to assess clinical information quality. *Stud. Health. Technol. Inform.,* 160(Pt 1). 699–703.

DataCleaner. (2014) *DataCleaner Monitor*. Available from http://datacleaner.org/ (12 June 2014).

DataFlux Corporation. (2003) *Data Profiling: The Foundation for Data Management*. Available from http://infoimpact.com/articles/Data%20Profiling%20White%20Paper1003-final.pdf (17 January 2014).

Eckerson, W. (2004) Data profiling: A tool worth buying (really!). *DM Rev.,* 14(6). 28-31,82.

Karr, A. F., Sanil, A. P. and Banks, D. L. (2006) Data quality: A statistical perspective. *Stat. Methodol., 3*(2), 137–173.

Liaw, S. T., Rahimi, A., Ray, P., Taggart, J., Dennis, S., de Lusignan, S., Jalaludin, B., Yeo, A. E. and Talaei-Khoei, A. (2012) Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *Int. J. Med. Inform.,* 82(1). 10-24.

Lorence, D. (2003) Measuring disparities in information capture timeliness across healthcare settings: effects on data quality. *J. Med. Syst.,* 27(5). 425-433.

Mansingh, A. and Srikant, S. (2005). Data profiling in a compliance world. *DM Rev*., 15(10). 52-56.

Miriyala, S. (2007) The need for data profiling in customer data integration. *DM Rev.,* 17(2), 6.

Mphatswe, W. W., Mate, K. S., Bennett, B. B., Ngidi, H. H., Reddy, J. J., Barker, P. M., and Rollins, N. N. (2012) Improving public health information: A data quality intervention in KwaZulu-Natal, South Africa. *Bull. World Health Organ.*, 90(3). 176-182.

Naumann, F. and Rolker, C. (2000) Assessment methods for information quality criteria. In: *Proceedings of the International Conference on Information Quality (IQ)*. Boston, MA.

Orfanidis, L., Bamidis, P. and Eaglestone, B. (2004) Data quality issues in electronic health records: An adaptation framework for the Greek Health System. *Health Inform. J.,* 10, 23-26.

Pipino, L., Lee, Y. W. and Wang, R. Y. (2002) Data quality assessment, *Commun. ACM*., 45(4). 211-218.

QlikView.(2014) http://www.qlik.com/us/explore/products/overview (12 June 2014).

Redman, T. C. (2001) *Data Quality: The Field Guide*. Digital Press: Boston, MA.

Shariat Panahy, P., Sidi, F., Affendey, L., Jabar, M. A., Ibrahim, H. and Mustapha, A. (2013) A methodology to explore rules and methods for data quality dimensions toward improvement the quality of databases. *J. Appl. Sci.,* 13(4). 615-620.

Strong, D. M., Lee, Y. W., and Wang, R. Y. (1997) Data quality in context. *Commun. ACM*., 40(5). 103-110.

Svensson-Ranallo, P. A., Adam, T. J. and Sainfort, F. (2011) A framework and standardized methodology for developing minimum clinical datasets. *AMIA Summits Transl. Sci. Proc.,* 2011. 54-58.

Wang, R. Y. (1998) A product perspective on total data quality management. *Commun. ACM*., 41(2). 58-65.

Wang, R. Y. and Strong, D. M. (1996) Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.,* 12(4). 5-33.

Weiskopf, N. G, and Weng, C. (2013) Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *JAMIA,* 20(1). 144-151.