Proceedings Paper:
Fox, F, Aggarwal, VR orcid.org/0000-0003-0838-9682, Whelton, H et al. (1 more author) (2018) A Data Quality Framework for Process Mining of Electronic Health Record Data. In: 2018 IEEE 6th International Conference on Healthcare Informatics (ICHI). IEEE ICHI 2018, 04-07 Jun 2018, New York, NY, USA. IEEE , pp. 12-21. ISBN 978-1-5386-5377-7

https://doi.org/10.1109/ICHI.2018.00009

A Data Quality Framework for Process Mining of Electronic Health Record Data

Frank Fox
School of Dentistry
University of Leeds
Leeds, U.K.
dnfgf@leeds.ac.uk

Vishal. R. Aggarwal
School of Dentistry
University of Leeds
Leeds, U.K.
V.R.K.Aggarwal@leeds.ac.uk

Helen Whelton
College of Medicine and Health
University College Cork
Cork, Ireland
h.whelton@ucc.ie

Owen Johnson
School of Computing
University of Leeds
Leeds, U.K.
O.A.Johnson@leeds.ac.uk

Abstract- Reliable research demands data of known quality. This can be very challenging for electronic health record (EHR) based research where data quality issues can be complex and often unknown. Emerging technologies such as process mining can reveal insights into how to improve care pathways but only if technological advances are matched by strategies and methods to improve data quality. The aim of this work was to develop a care pathway data quality framework (CP-DQF) to identify, manage and mitigate EHR data quality in the context of process mining, using dental EHRs as an example.
Objectives: To: 1) Design a framework implementable within our e-health record research environments; 2) Scale it to further dimensions and sources; 3) Run code to mark the data; 4) Mitigate issues and provide an audit trail.
Methods: We reviewed the existing literature covering data quality frameworks for process mining and for data mining of EHRs and constructed a unified data quality framework that met the requirements of both. We applied the framework to a practical case study mining primary care dental pathways from an EHR covering 41 dental clinics and 231,760 patients in the Republic of Ireland.
Results: Applying the framework helped identify many potential data quality issues and mark-up every data point affected. This enabled systematic assessment of the data quality issues relevant to mining care pathways.
Conclusion:
The complexity of data quality in an EHR-data research environment was addressed through a re-usable and comprehensible framework that met the needs of our case study. This structured approach saved time and brought rigor to the management and mitigation of data quality issues. The resulting metadata is being used within cohort selection, experiment and process mining software so that our research with this data is based on data of known quality. Our framework is a useful starting point for process mining researchers to address EHR data quality concerns.

Keywords – EHR, research data, process mining, data quality

I. INTRODUCTION

Electronic Health Record (EHR) systems are now well established in many countries and healthcare settings. The importance of the secondary use of EHR data for research is widely recognized. Reliable research demands data of good quality or, at least, data of a known quality and without this, research results are impossible to evaluate. Robust data provenance and data of acceptable and known quality must become the norm.

There is a growing body of literature that uses data derived from EHRs to inform medical and health research. There is also a growing, but noticeably smaller, body of literature on the underlying data quality (DQ) problems inherent in EHRs as a research data source which address the huge scope for non-random human error across multiple dimensions. Frameworks such as those proposed by Weiskopf and Weng [1] and Kahn et al. [2] can be used to categorize the dimensions of EHR DQ and help identify suitable strategies for mitigation. Their adoption in EHR research is urgent.

Process mining is a set of emerging tools and techniques that can be used with EHR data to examine temporal patterns of care provision including mining, modeling and measuring patients' experience of care pathways. Event logs are extracts from EHRs comprising of lists of time-stamped process-steps created as a by-product of operations. Process mining tools use these widely-available event logs to produce visualizations of the real-world processes that EHRs support. The approach can generate unique insights on process execution, resource usage and conformance. Their analysis has the potential to identify bottlenecks, causes of delayed diagnosis and the optimum pathways to support precision medicine. These processes might be clinical pathways or business processes within or across organizations.

Process mining in healthcare is challenging because patterns of care vary widely between patients, between health care professionals and organizations and the reliance of the method on the completeness of time-stamped event logs adds additional requirements for measurable DQ. As with other forms of data mining, systematic logging, repair and analysis techniques are important, as is the need for transparency around data cleaning and checking steps. This paper presents a Data Quality Framework that addresses the specific needs of process mining of Care Pathways (CP-DQF) based on the process mining and EHR DQ literature. The framework enables DQ to be managed systematically to support more reliable process mining work in EHR research.

DQ issues can arise at any time in the lifetime of EHR data, from the design of the underlying EHR application and database, its use in practice, through to the extraction of data for research and the technologies and methods used there. How does the CP-DQF help with this complex problem? The CP-DQF framework helps

i) identify DQ issues
ii) record DQ issues
iii) mark-up research datasets with DQ metadata
iv) mitigate effects of DQ issues on research by easing exclusion of data
v) mitigate effects of DQ issues on research by imputation of values or other methods
vi) report on the extent and impact of DQ issues.

We applied the framework to a practical case study mining primary care dental pathways from a large dentistry EHR covering 41 dental clinics and 231,760 patients in the Republic of Ireland.

In the following sections, we review the background to EHRs and process mining of care pathways identifying specific DQ dimensions and information sources. We describe the CP-DQF, its strategy, structure and application and illustrate its use in the case study before reflecting on further work. While this paper is focused on process mining, we believe that the framework provides a new and structured way of approaching EHR DQ applicable to the wider health informatics community.

II. Background

The use of EHR data for healthcare research is gathering momentum and is supported by business [3], health authorities and governments [4] [5] [6].

Many benefits from using EHR data for research have been identified including epidemiology, disease outcomes, pharmaco-vigilance and comparative effectiveness [7]. Other uses include syndromic surveillance, public health, research and quality improvement [8] [9]. Rapid cohort identification, quality of care assessment, research, data privacy and de/re-identification have also been identified as areas where access to clinical data can aid researchers [10]. More recently, there has been growing interest in process mining in healthcare [11] [12], including specialist healthcare areas such as stroke-care [13], diabetes [14] and oncology [15]. Process mining uses event logs originating from traces left by the execution of the real-world process in the organization's information systems and are a by-product of the organization's operations. Process mining can add value to established data-mining methods by producing visualizations of real-world processes including care pathways and patterns of care. It is seen as bridging the gap between traditional model-based process analysis and data-centric analytics or data mining and although it is still an emerging discipline, it has much to offer for health informatics systems audiences [16]. An example of process visualization of a dental care pathway is shown in Figure 1 below.

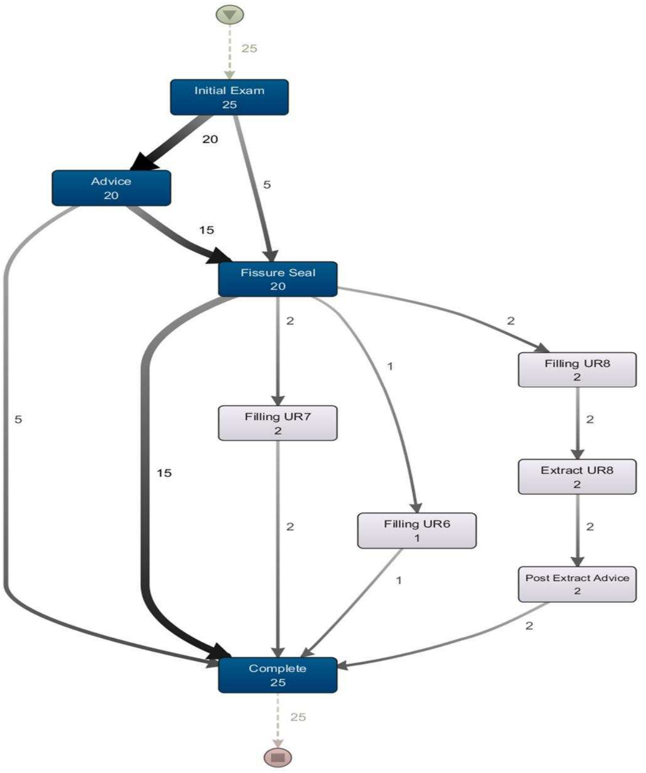


Figure 1: Process visualization of a dental care pathway from our dataset

The visualization was produced by creating and transforming an extract of time-stamped activities from our dental care EHR (described below) and loading into the DISCO process mining tool (www.fluxicon.com/disco). The graph illustrates typical primary care dental consultations starting with the event 'Initial Exam' and the various paths that are typically followed by patients, through to the common last event, 'Completed Case'. Because the pathway is mined from real EHR data the actual number of patients experiencing each event can be shown, for example, of the 25 patients examined 20 received the preventive measure, Fissure Seal. More common paths and more frequent activities are denoted by thicker arcs and darker boxes respectively.

In data and process mining of EHRs, DQ issues are complex and can arise from many sources. DQ can also affect the data at different levels, arise at different times in the research and can come from different root causes. The secondary use of EHR data for research demands validated, systematic methods of DQ assessment [1] and there is correspondingly urgent need for process mining to incorporate techniques addressing DQ problems [17]. Most authors encourage systematic logging techniques and the development of repair and analysis techniques with the objective of improving the quality of the event logs and consequently, improving the outputs of process mining exercises. Greater transparency around data cleaning and checking steps is also advised [1]. DQ frameworks should support the discovery, management and mitigation of these issues. DQ issues should be logged in a systematic fashion and carefully documented with an audit trail that includes evidence, assumptions and decisions.

Researchers need a framework that bridges the gap between process mining and DQ and is implementable as a software tool. This suggests an environment for describing the DQ issues and logging their impact, marking the data and applying mitigation strategies. Such a framework should ensure data of known quality underpins research that uses process mining to improve care pathways.

A. Dimensions of DQ.

Dimensions of DQ allow us to identify data features we can measure. Our starting point was the review of the literature on dimensions of EHR DQ and methods of DQ assessment by Weiskopf & Weng [1]. Completeness, correctness, concordance, plausibility and currency were identified as the dimensions and seven broad categories of assessment methods were also identified. They further suggested that concordance and plausibility could be handled within the 'correctness' dimension. Many other dimensions were identified in this review which they rationalized to the five named above. This framework has been successfully applied to MIMIC-III, a publicly available e-health record database [18]. Incomplete or missing data, inconsistent and inaccurate data are confirmed as major issues [7] [8] [9] [10]. A variation of these dimensions is also proposed [19] [20]: Completeness, conformity, consistency, accuracy, validity and duplication. Kahn et al. [2] produced a harmonized DQ assessment terminology and framework for the secondary use of EHR data incorporating several existing EHR DQ frameworks. Their output consisted of harmonized DQ terms and an organizing framework. They further rationalized DQ dimensions into 3 categories; Conformance with subcategories value, relational and computational, Completeness, and Plausibility with subcategories uniqueness, atemporal and temporal. These categories can be applied in two assessment contexts; Verification (internal to the data) and Validation (referencing external benchmarks). Intrinsic data features were included in the scope of the study with extrinsic features including fitness for a specific analysis excluded. DQ issues caused by deficiencies in the data representation or the data model and 'relevancy' were also excluded.

Specific to process-mining, four broad DQ issues that could exist in process mining event logs were identified by [11] [17]: missing, incorrect, imprecise and irrelevant. This further dimension, 'irrelevant', is very interesting to process miners because superfluous information increases the complexity of process-maps and detracts from their comprehensibility. These dimensions were further detailed in 27 types of quality issues relating to the case, event and attribute levels of the data in an event log. The widely cited Process Mining Manifesto proposes a rating system for DQ ranging from 1-star to 5-star [21].

The proposed framework allows us to include and to tailor those dimensions and categories appropriate for the specific research and to include extrinsic data features as DQ issues.

B. Where are the DQ information sources?

Assessing EHR DQ is complicated by the many potential sources of information. Having established the dimensions of DQ relevant to the research, we now need to establish specific DQ issues. Some of the many stakeholders with potentially valuable commentary on DQ are summarized below:

- Software developers and database administrators
- EHR application users
- Domain experts
- Previous research work using this data or similar data
- General EHR DQ literature
- Technology specific literature (Process Mining)
- Comparison to standards (HL7, SNOMED-CT, ANSI etc.).

III. Method: The Care Pathway Data Quality Framework (CP-DQF)

A. Introduction

Applying Deming's Plan-Do-Study-Act [22] our overall approach for using the CP-DQF is:

- Plan – Frame the quality questions for the research
- Do - Identify DQ dimensions. Identify potential sources of information on DQ. List potential DQ issues. Relate the issues to the experiments. Mark the data. Mitigate the DQ issue if possible
- Study – Analyse the results of the 'Do' phase
- Act – Take steps to improve future DQ.

The aim here is that data of unacceptable quality is marked as 'bad' i.e. unusable. Imperfect but acceptable data is marked as 'compromised' i.e. it can be used in some situations or experiments. The remaining data is unmarked or 'good' and is available for all purposes.

The framework can incorporate fitness-for-use DQ issues, i.e. DQ issues affecting specific experiments. Involvement of the researchers or principle investigators at this juncture will strengthen the exercise and help eliminate confounders and invalid assumptions. The CP-DQF maintains a registry of DQ issues. Code is written to mark individual data elements (usually rows) affected by the DQ issue. The code is stored with the DQ issue in the registry. In the case-study below, this code consisted of Structured Query Language (SQL) update commands. The research data is assessed against the DQ registry and data records affected by these issues are marked. The scale of each issue is recorded and mitigated through code if possible. The registry is currently managed with manual inserts and updates and building a user interface is in progress. The principle components of the data structure supporting the CP-DQF are shown in the entity relationship in Figure 2.

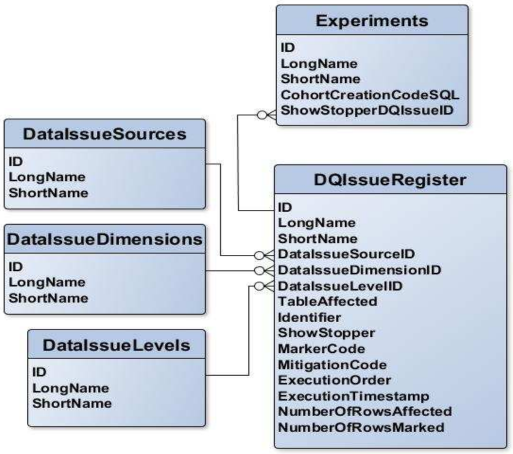


Figure 2: CP-DQF Entity-Relationship Diagram

B. Using the CP-DQF.

Using the CP-DQF has three main steps. First, establish the DQ issues register for the research. Known issues are pre-populated in the register and this is supplemented with additional issues specific to the research questions. Second, push the research data through the CP-DQF analysis tool. Third, report on what happened.

Step 1: Establishing the DQ issues register for the research.

- Add general DQ issues to the register
- Create entries in the `Experiments` table
- Add any experiment-specific issues to the register
- Link `Experiments` to entries in the `DQIssuesRegister.` This will disqualify the data from use in that experiment if marked as a showstopper.

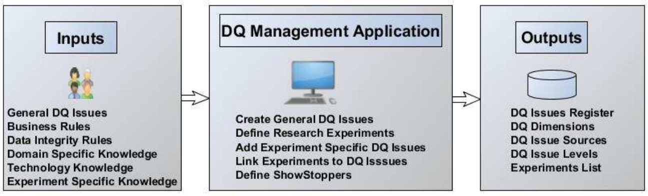


Figure 3: CP-DQF (Step 1)

The outputs from Step 1 consist of three dictionary entities: `Dimension, Levels` and `Sources`, as well as the DQ issues register itself and a list of experiments for this research.

Step 2: Applying the CP-DQF to the research data.

A number of steps are taken in applying the CP-DQF to the research data.

1) Add Metadata to the research data.

Mark-up fields are added to the research data allowing us to store DQ information with the data element (usually a row). This information can be used to exclude the data from the dataset as it is extracted for a specific experiment. Suggested fields are: a Boolean called `BadRow` and a vector string called `BadRowCodes`. The vector string can hold multiple error codes simultaneously.

2) Pre-processing or discussion section?

Decide where the DQ issue is to be dealt with, in pre-processing or by way of discussion. This will determine whether we can mark the data with this issue or not. If not, we will address it in the research discussion.

3) Does an issue disqualify the data from the experiment?

If the DQ issue is serious for any specific experiment, the experiment should be marked, and the data excluded from use there.

4) Evaluate the effect of these data disqualifications.

Does it require re-execution of previously executed marking or mitigation code? Does it skew results? E.g. Removal of data may violate previously satisfied data integrity constraints.

5) Write/Run the Marking Code from the CP-DQF against the data.

Executing the code stored with the DQ issue in the register will mark the research data's metadata with information about its DQ.

e.g. Mark orphaned treatments (no client exists) as 'bad'

```
Update PMTreatments
set BadRow =1,
BadRowCodes= Concat(BadRowCodes,’ 7’)
where ClientID not in (select PMClientID
from PMClients);
```

6) Write/Run the Mitigation Code against the data.

Executing the mitigation code (if exists) will update the research data to improve its quality.

7) Update the DQ issues register with the results.

Record the scope of the issue and the scope of the mitigation efforts – primarily for reporting purposes.

8) Write/Run the CohortSelection Code.

Cohort/Dataset selection code can now be written incorporating the metadata as a criterion for exclusion/inclusion in the data set. In the implementation below, treatment events are only selected if the metadata, `BadRow` is NULL.

e.g.
```
Select * from PMTreatments
where ClientAge = 8
and   BadRow is NULL
```

These steps are summarized in Figure 4

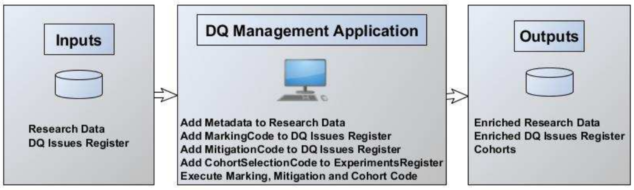

Inputs
Research Data
DQ Issues Register
DQ Management Application
Add Metadata to Research Data
Add MarkingCode to DQ Issues Register
Add MitigationCode to DQ Issues Register
Add CohortSelectionCode to ExperimentsRegister
Execute Marking, Mitigation and Cohort Code
Outputs
Enriched Research Data
Enriched DQ Issues Register
Cohorts

Figure 4: CP-DQF (Step 2)

C. Step 3: Analysis and Report on Steps 1 & 2.

- Report on the DQ; how much data was affected etc.
- Evaluate the effect of data disqualifications in Step 2 above. e.g. do the changes change or bias the results?

D. User Interface/ Infrastructure.

CP-DQF currently uses scripting to create registry entries, mark the code and report on the impacts. A user interface is being developed. Deployment in other databases is being considered. Error handing has been considered. Client/Server architecture has not been considered. Data/business rules/UI layers have not been considered.

IV. Results: Validating the CP-DQF

A. Introduction.

The data used to validate the CP-DQF was an extract from a single-center relational EHR database containing information relating to patients and their dental treatment under the Health Service Executive (South), Ireland, covering 41 clinics in two counties, Cork & Kerry. All these clinics used the same Bridges Software EHR system with a centralized database and common database schema. Our research data extract covered the public health dental screenings of school children attending these clinics (n=231,760) between 2000 and 2014 and includes clinical charts (n=1,016,197), treatment events (n=3,169,864) and tooth conditions (32,291,681). The information technology environment used was Microsoft SQL Server 2017 Server Management Studio with DISCO and ProM (www.promtools.org) being used for process mining.

B. Data Acquisition Process.

Ethical approval was granted by the Clinical Research Ethics Committee of the Cork University Teaching Hospitals (CREC) on August 2nd, 2016 (Reference: OHSRC00516) and subsequently permission was granted by the HSE Primary Care Research Committee (PCRC), Republic of Ireland at its meeting on January 17th, 2017.

C. Validation - Stage 1 Establish the DQ issues register for this data.

To identify the potential DQ issues the following steps were taken:

1) Identify potential DQ information dimensions.

While cognisant of the quality dimensions proposed in [1], [2], [19] & [20] we used those proposed in [11], designed specifically for Process Mining.

- Incomplete (e.g. missing date-of-birth)
- Incorrect (e.g. incorrectly logged timestamp)
- Imprecise (e.g. lacking precision or too coarse)
- Irrelevant (e.g. increasing complexity of process map without contributing value)

We mapped some of the other proposed dimensions to the above: 'concordance' and 'plausibility' to 'incorrect' [1];

‘accuracy’ to ‘imprecise’ [19] and ‘conformity, consistency, validity and duplication’ [19] to ‘incorrect’, ‘conformance’ and ‘plausibility’ [2] to ‘incorrect’.

2) Identify DQ information sources in these dimensions. The DQ information sources for this research are listed in Section II (B).

3) List potential DQ issues from these sources

a) From the software developers/ database administrator.

The software developer/ database administrator identified a set of data integrity rules aimed primarily at marking data in the data extract that should not have been there in the first place. In the main, this related to orphaned data. In the dataset, 23 such integrity and business rules were identified. Each rule has an entry in the DQ registry and code was written to update each affected record affected by the rules.

b) From the application users.

[23] described how day-to-day users of the Bridges EHR system were studied, and analysis and verification of data entry practice was carried out. Recording of gender, fluoridation status and dental trauma were identified as invalid due primarily to incorrect data entry protocols. These attributes have been excluded from the research herein. As an example, the event named ‘Initial Exam’ was intended to represent a school screening however the research revealed the code was being used in other circumstances with a 50% compliance being established [23].

c) From Dental Domain Experts.

This was not carried out in the DQ assessment phase of this research and may be more appropriate to the discussion section. No issues from this source are entered in the registry at this time. This area offers strong potential for calculated metrics such as mean, median, and value distributions. Validated oral health benchmark measures such as DMFT (Decayed, Missing & Filled teeth) could be registered here and the research data values compared to this to give an indication of external data validity. Other work, specific to the implementation of dental quality measures in dental EHRs [24] should also provide indications of external data validity.

d) From earlier research using this data.

As in the ‘From the application users’ section above, gender, fluoridation status and dental trauma status were not used, and these DQ issues can be ignored in this research. There were no additional issues from this source entered in the DQ register.

e) General Data Mining Literature.

The general data mining literature suggests common issues are representational bias, clinician-related biases regarding missing data and outcomes, non-standardization of data entry, data redundancy, inaccuracy, restriction to retrospective study, and difficulties extracting data [7]. Root causes for some DQ issues in the secondary use of data created for project management of EHR implementation were identified as: Differential incentives for the recording of data i.e. data tended to be more accurately recorded if needed for contractual or financial purposes; flexibility in the software systems allowing multiple ways of doing the same task; variability in documentation practices between personnel; variability in the use of standardized vocabulary and changes in procedures and electronic system configuration over time [8].

Missing, inaccurate and inconsistent data were also identified as issues in a study of pancreatic cancer data [9] and attributed to information fragmentation in the healthcare system and poor documentation of critical information. Inaccuracies were also caused by poor granularity of diagnosis terms or incorrect us of the terms. Inconsistencies arose due to different data sources in the EHR and inconsistent use between clinicians. The authors also proposed some solutions involving formal information exchange mechanisms, clinical registries and personal health records as well as the sharing of effective strategies for secondary use of healthcare data [9].

f) From Process Mining DQ Literature.

Four broad DQ issues that exist in process mining event logs were identified by [11] [17]: Missing Data, Incorrect Data, Imprecise Data and Irrelevant Data. These were further detailed in 27 types of quality issues.

The research dataset extract was evaluated for each of these, identifying whether it is likely that the problem exists, how it may have arisen and what its effect is likely to be. Further, we considered steps to mitigate the problem and whether their effect merits the investment. Using the method proposed in [11] we tabulated possible sources of DQ issues. Potential issues were numbered as in [11] with ‘N’ indicating that the issue does not exist, ‘L’ indicates a low likelihood of the issue being present and ‘H’ indicating a high likelihood summarized in Table1.

Table 1: 27 DQ Issues adapted from Mans, et al., (2015)

	Missing	Incorrect	Imprecise	Irrelevant
Case	1 (L)	10 (L)	N/A	26 (L)
Event	2 (L)	11 (L)	N/A	27 (L)
Relationship (Belongs to)	3 (N)	12 (N)	19 (N)	N/A
C_attribute	4 (N)	13 (N)	20 (N)	N/A
Position	5 (L)	14 (H)	21 N/A	N/A
Activity Name	6 (N)	15 (L)	22 (L)	N/A
Timestamp	7 (N)	16 (N)	23 (N)	N/A
Resource	8 (H)	17 (L)	24 (L)	N/A
E_attribute	9 (H)	18 (L)	25 (N)	N/A

Each of these 27 issues was entered in the registry for completeness. Additionally, categories of process characteristics were proposed in [17] with the potential to impact the output of process mining; voluminous data, case heterogeneity, event granularity and process flexibility.

g) From Standards.

Care pathways are often highly variable in clinical settings and process mining of EHRs often produce logs of high heterogeneity and very fine granularity leading to spaghetti–like process models, so-called because the process maps produced are so complex they appear similar to a plate of spaghetti. To untangle the spaghetti, abstraction methods using classifiers or ontologies are commonly used, for example, abstractions or standards like SNODENT-CT, [25]. Trace clustering has been shown to be effective in identifying patients with similar pathways, which can be then be used to partition event logs into subsets of homogeneous cases.

4) Create entries in the Experiments table.

In the validation phase, we set up just one experiment comparing two cohorts at age 13/14 - one which had received school dental screenings and fissure-sealants at age 8, and one receiving no fissure-sealants. Cohorts were assessed using the decayed, missing and filled teeth (DMFT) index; the data was not adjusted for factors that can confound DMFT.

5) Add experiment-specific DQ issues to the register.

None arise.

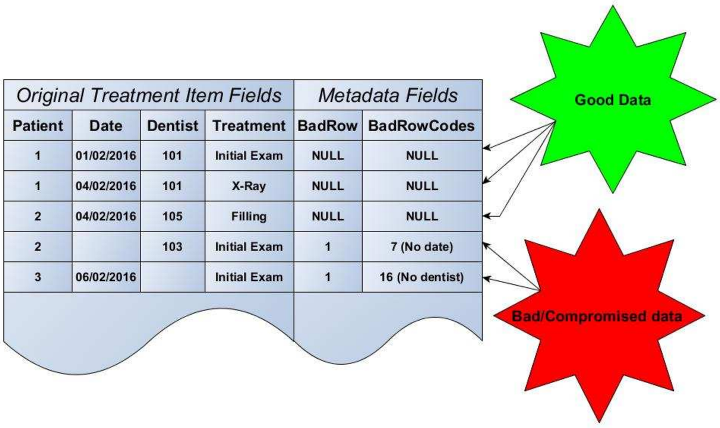
6) Identify DQ 'Showstoppers' and mark experiments with them.

The experiment's showstopper is marked to indicate that there is a showstopper entry in the `DQIssuesRegister`. This means that data marked with this DQ issue will be excluded from the research. The DQ issue is, 'All entries in `PMTreatments` must have a corresponding `Client` in `PMClients`'.

D. Validation - Phase 2: Applying the CP-DQF to the research data.

1) Add Metadata to the research data.

In this implementation, two additional fields were added, a Boolean called `BadRow` and a vector string called `BadRowCodes`. The vector string can hold multiple appended error codes if required. This structure is represented in Figure 5 below.



Original Treatment Item Fields				*Metadata Fields*	
Patient	Date	Dentist	Treatment	BadRow	BadRowCodes
1	01/02/2016	101	Initial Exam	NULL	NULL
1	04/02/2016	101	X-Ray	NULL	NULL
2	04/02/2016	105	Filling	NULL	NULL
2		103	Initial Exam	1	7 (No date)
3	06/02/2016		Initial Exam	1	16 (No dentist)

Figure 5: Example of Research Data with Metadata added

2) Pre-processing or discussion section.

The DQ issue here is a data integrity issue and accordingly can be dealt with here in the DQ pre-processing section.

3) Decide which of these issues disqualifies the data from use in the experiment.

If the DQ issue is serious for any specific experiment, the experiment should be marked, and the data excluded from use. In this case, all treatments should have a valid active client. If there is no client associated with a treatment, vital information is missing e.g. the age of the client. Therefore, this issue disqualifies the data from use in this age-dependent experiment.

4) Evaluate the effect of these data disqualifications.
Does it require re-execution of previously executed marking or mitigation code. Does it skew results? Depending on the extent of the issue and the underlying causes, this may cause skewing of the data.

5) Write/Run the Marking Code from the CP-DQF against the data.
Executing the code stored with the DQ issue in the register marked the research data's metadata with information about its DQ.

6) Write/Run the Mitigation Code from the CP-DQF against the data.
No mitigation code is applied directly to the data at this point. However, the fact that the data is now annotated with DQ information allows exclusion of specified data from individual experiments which is intended to have the effect of mitigating the DQ issue. Code to directly mitigate the DQ issue e.g. imputation of missing values is being developed.

7) Update DQ issues register with the results.
Here, the scope of the DQ issue and the scope of the mitigation efforts are recorded are added to the `DQIssuesRegister` for reporting purposes.

8) Write/Run the CohortSelection Code.
Cohort/Dataset selection code can be executed incorporating the metadata as a criterion for exclusion/inclusion in the data set. In our implementation, treatment events are only selected if the metadata, `BadRow,` is NULL.

E. Step 3: Report on Step 1 & 2.

After executing steps 1 & 2, it is important to know the scope of the DQ issues and to this end a report showing DQ metrics can be run against the `DQIssuesRegister`. The report should list the issues in the register along with frequency and percentage data affected. This may flag issues needing attention and a root cause analysis might be needed leading to improvement steps and better future DQ. The predominant metric used shows a 'percentage' indicating the scale of the DQ issue against the total number of rows. Practically, this only applies to DQ issues at the row or field level. Other metrics, e.g. those comparing calculated values such as mean, median and distributions to expected values are also calculated at this step. A sample of the data issues in the registry and their scope is shown in Table 2 below.

Table 2: Example of `DQIssuesRegister` entries

DQ Issue Name (Integrity Rules)	**Rows marked**	**% Affected**
All entries in `PMTreatments` must have a corresponding Client in `PMClients`	48330	1.52
All entries in `PMChart` must have a corresponding Client in `PMClients`	3267	0.32
All entries in `PMQuestionnaire` must have a corresponding Client in `PMClients`	1813	0.55
All entries in `PMQuestionAnswers` must have a corresponding Client in `PMClients`	50806	0.52
All entries in the `PMTreatments` must have a `MappedToProcedureNameGroup` in the `PMProcedureGroupNames` table - to reduce noise from rarely occurring procedures (<100 times)	18316	0.57
All Treatments in the `PMTreatments` must have a `CompletionDate` >=1990-01-01 00:19:02.000	197352	6.22

V. LIMITATIONS & FUTURE WORK

- Our case-study describes a scenario where the researcher has direct access to the data through SQL Server Management Studio. This access allowed addition of the metadata fields to the research data, database scripting, inclusion of additional clauses in the cohort selection process etc. The current framework design incorporates assumptions based on this scenario. Different research scenarios may require alternative approaches, for example, storing the DQ metadata in distinct and separate tables or locations, or database normalization measures.
- The proposed database design (Figure 2) fulfills the requirements of the case-study herein. Other scenarios may require redesign. Simpler case-studies may only require the `DQIssuesRegister` while more complex scenarios may require further normalization of the database to improve data integrity and reduce data redundancy. It is unknown how this would impact the performance of cohort selection queries.
- Our case-study deals with research data from a single, homogeneous EHR source. Consideration needs to be given to additional DQ matters such as 'Variety' in scenarios with complex, multi-source, multi-institution research projects using heterogeneous data sources - perhaps as approached by Knowlton et al [26].
- Our case-study is based on process mining and used the DQ dimensions from Mans et al [11]; Incomplete, Incorrect, Imprecise and Irrelevant. Further work to incorporate the dimensions from Kahn et al [2] and others could contribute to a more harmonized and generalizable understanding. The

CP-DQF framework is customizable allowing the incorporation of these additional DQ dimension, however, the deeper thinking behind these dimensions must the reconciled with the requirements of process mining research work to avoid overlap of dimensions and gaps. In particular, the important extrinsic data features such as ‘fitness-for-use’ and ‘relevancy’, which are central to our process mining research need to be included in the framework.

- The design presented here could be developed to further encompass data management in research using EHR data. This might include logging and auditing other elements of the Extract, Transform, Load process, multiple runs of the same experiment, user management and error handling etc.
- While some of the DQ issues can be identified, marked and perhaps mitigated-against in a pre-processing phase of the research e.g. Missing Date-of- Birth, others are less clear-cut, and might only be adequately dealt with by way of discussion e.g. issues caused by clinician bias, researcher bias, or data model deficiencies. The distinguishing line between these types of issues is unclear to us and would benefit from further work. It seems likely that many of these types of issues may be difficult or impossible to automatically identify and mitigating these issues may be multi-faceted and require root-cause analysis.
- Future work can include approaches from latent class imputation to mitigate missing data.
- The results presented have focused on a small number of easily quantifiable DQ issues with the easily established metric of ‘% affected’. More complex DQ metrics as detailed above are in development.
- Further metrics could also be added to the data based on the method of DQ assessment employed [1] e.g. gold-standard assessment methods would give the overall DQ a higher rating.
- Assessing whether exclusion of the quality-affected data impacts the outcomes of specific research experiments would be useful.
- Specific and detailed questions on DQ could be developed and embedded within the live EHR e.g. To the Application Users - “Is there any possibility that Date of Birth has been incorrectly recorded?”

VI. Conclusions

The design for the CP-DQF data quality framework has been presented. It is implementable as a software tool that can be used to manage the DQ issues of research using EHRs. In our work we have applied the CP-DQF framework to a large dental EHR and the framework proved useful in providing a structured method to identify and document issues following the DQ dimensions established by the existing literature, notably [1] [11]. Our example illustrates how code to mark the data to mitigate DQ can be implemented. Intimacy with the data was helpful in identifying many of the information sources and data issues. The case study also showed DQ issues linked to individual experiments in the research and how this can cause affected data to be excluded if appropriate. The CP-DQF framework has the functionality to be used as an audit trail tool for all data transformations and data cleaning activities. This would satisfy the demands for greater transparency in the pre-processing of EHR-data in preparation for research. By slightly varying the cohort selection criteria, it is also possible to compare research results before and after the exclusion of bad quality data to determine whether its impact. While the framework was prototyped in the Microsoft SQL Server environment, researchers in other environments could easily replicate this design. The entity design is simple but effective and the dictionaries of sources, dimensions and levels can be tailored to the research.

Use of the CP-DQF may help researchers think about the potential DQ issues in their research, log and manage them in a structured environment, create an audit trail for data transformations, assess and mark their data with quality information, mitigate the issue if possible, exclude data from their experiments if appropriate, compare before and after research outputs and finally, report on DQ metrics.
This will lead to known and more robust EHR DQ, a secure audit trail of DQ transformations, reproducible research steps and more reliable process mining results.

Research conclusions can and should be informed by a rigorous assessment of DQ and a structured and auditable approach to marking and mitigating DQ issues. Our framework provides a useful starting point for other process mining researchers to address EHR DQ concerns.

VII. References

[1] N. 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, vol. 20, no. 1, pp. 144-151, 2013.

[2] M. Kahn, T. Callahan, J. Barnard, A. Bauck, J. Brown, B. Davidson, H. Estiri, C. Goerg, E. Holve, S. Johnson, S. Liaw, M. Hamilton-Lopez, D. Meeker, T. Ong, P. Ryan, N. Shang, N. Weiskopf, C. Weng, M. Zozus and L. Schilling, "A harmonised data quality assessment terminology and framework for the secondary use of Electronic Health Record Data," eGEMs (Generating Evidence & Methods to improve patient outcomes), vol. 4, no. 1, 2016.

[3] McKinsey Global Institute, "Big Data: The next frontier for innovation, competition and productivity," 2011. [Online]. Available: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation. [Accessed 19 July 2017].

[4] The Parliamentary Office of Science and Technology, "Big Data and Public Health (POSTNOTE 474)," 19 July 2017. [Online]. Available: http://researchbriefings.parliament.uk/ResearchBriefing/Summary/POST-PN-474.

[5] M. Wilson, N. Dumontier, I. Jan Aalbersberg, G. Appleton, M. Axton, A. B. N. Baak and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, pp. 1-9, 2016.

[6] European Commission, "Research and Innovation Performance in the EU," Publications Office of the European Union, Luxembourg, 2014.

[7] M. Song, L. Kaihong, R. Abromitis and T. Schleyer, "Reusing Electronic Patient Data for Dental Clinical Research: A Review of Current Staus," Journal of Dentisry, vol. 41, no. 12, pp. 1148-1163, 2013.

[8] I. Danciu, J. Cowan, M. Basford, X. Wang, A. Saip, S. Osgood, J. Shirley-Rics, J. Kirby and P. Harris, "Secondary use of clinical data: The Vanderbilt approach," Journal of Biomedical Informatics, pp. 28-35, 2014.

[9] J. Anker, S. Shih, M. Singh, A. Snyder and A. K. R. Edwards, "Root Causes Underlying Challenges to Secondary Use of Data," 2011.

[10] T. Botsis, G. C. F. Hartvigsen and C. Weng, "Secondary Use of EHR: Data Quality Issues and Informatics Opportunities," 2010.

[11] R. S. Mans, v. d. A. W. M. P. and R. J. Vanwersch, Process Mining in Healthcare. Evaluating and exploiting operational healthcare processes, Springer-Verlag, 2015.

[12] W. Rojas, J. S. M. Munoz-Gama and D. Capurro, "Process Mining in Healthcare: A Literature Review," Journal of Medical Bioinformatics, vol. 61, pp. 224-236, 2016.

[13] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini and W. van der Aalst, "Process Mining Techniques: an Application to stroke care," Studies in Health Technology and informatics, vol. 136, pp. 573-578, 2008.

[14] C. Fernandez-LLatas, A. Martinez-Milanna, A. B. J. Martinez-Romero and V. Traver, "Diabetes care related process modelling using Process Mining techniques," 2015.

[15] A. Kurniati, O. Johnson, D. Hogg and G. Hall, "Process Mining in Oncology: a Literature Review," Hatfield, UK, 2016.

[16] M. Thiede and D. Fuerstenau, "The technological maturity of Process Mining," Berlin, 2016.

[17] J. Bose, R. Mans and W. van der Aalst, "Wanna Improve Process Mining Results? It's High Time We Consider Data Quality Issues Seriously," BPM Centre Report BPM-13-02 BPMcentre.org, 2013.

[18] A. Kurniati, E. Rojas, D. Hogg and O. Johnson, "The Assessment of Data Quality Issues for Process Mining in Healthcare Using Mimic-III, a Publicly Available e-Health Record Database," Health Informatics Journal (Submitted for Publication), 2018.

[19] DAMA UK Working Group, "The six primary dimensions for data quality assessment," 2013. [Online]. Available: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf. [Accessed 30 1 2018].

[20] Microsoft, "SQL Server Data Quality Dimensions," 2012. [Online]. Available: https://social.technet.microsoft.com/wiki/contents/articles/3919.data-quality-services-dqs-faq.aspx. [Accessed 30 Jan 2018].

[21] IEEE, "Process Mining Manifesto," in Lecture Notes in Business Information Processing, Berlin, Heidelberg, Springer, 2011.

[22] R. Moen, "Foundation and history of the PDSA cycle," Feb 2010. [Online]. Available: https://s3.amazonaws.com/wedi/www/FileManager/PDSA_History_Ron_Moen.pdf.

[23] E. Murphy, "Can BRIDGES Dental Informatics System play a role in the development of an Evidence-based HSE Public Dental Service (MDPH Dissertation)," Department of Oral Health and Development, National University of Ireland, Cork, 2011.

[24] A. Bhardwaj, R. Ramoni, E. Kalenderian, A. Neumann, N. Hebballi, J. While, L. McClellan and M. Walji, "Measuring up. IMplementing a dental quality measure in teh electronic health record context," JADA, vol. 147, no. 1, pp. 35-40, 2016.

[25] C. Pedrinaci and J. Dominque, "Towards an Ontology for Process Monitoring and Mining," Semantic Business Process and Product Lifecycle Management, vol. 251, pp. 76-87, 2007.

[26] J. Knowlton, T. Belnap, B. Patelesio, E. Priest, F. Recklinghausen and A. Taenzer, "A Framework for Aligning Data from Multiple Institutions to Conduct Meaningful Analytics," eGEMs (Generating Evidence & Methods to improve patient outcomes, vol. 5, no. 5, pp. 1-7, 2017.