UNIVERSITY OF LEEDS

A Framework for Tackling Data Quality in Process-Oriented Data Science Research using Electronic Health Records.

Authors: F.Fox[1], V.R.Aggarwal[1], H.Whelton[2], O.Johnson[1], [1]University of Leeds, U.K., [2]University College Cork, Ireland

International Workshop on Process-Oriented Data Science for Healthcare (PODS4H) – Success Case

Research using data from electronic health records (EHRs) is in its infancy when compared to research using traditional methods for clinical trials and epidemiological studies.

Researchers must develop a detailed understanding of the EHR data's provenance, quality, and suitability before they can trust the data enough to answer the research questions being asked.

A key step to achieving that trust is to face the challenges of EHR data quality (DQ) head on: Find and document the issues, manage them, assess their impact and relevance to the research, mitigate their effects where possible, and report clearly on these steps.

We have developed a data quality framework[3] to achieve these aims and applied this to a Dental EHR-based process-mining research project in the University of Leeds.

The framework is based on existing EHR and process mining data quality literature, and is implementable as an automated, software solution. [1][2]

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] 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.

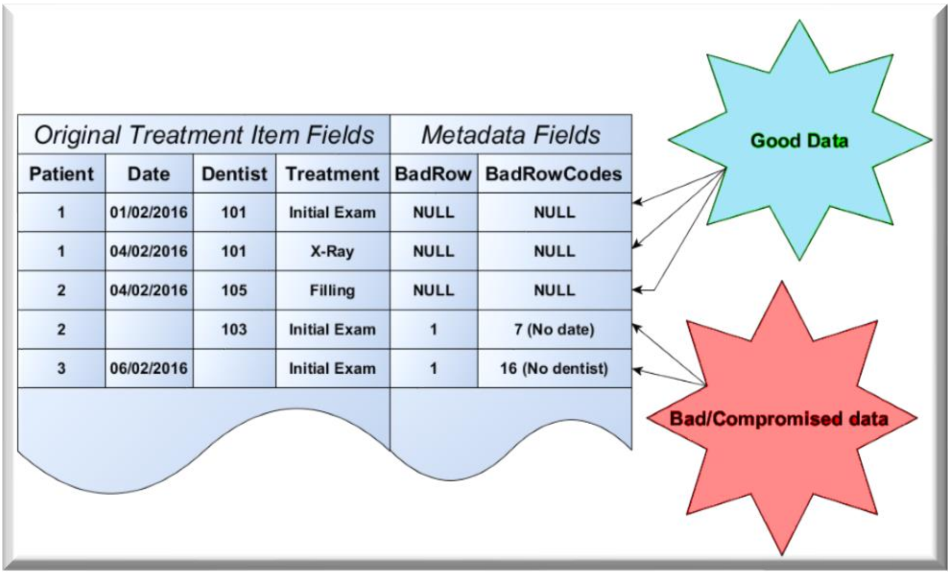
[3] F.Fox, V.R.Aggarwal, H.Whelton, O.Johnson, "A Data Quality Framework for Process Mining of Electronic Health Record Data", IEEE International Conference on Healthcare Informatics Proceedings, P12-21, New York, 2018.

> *"When discussing data and process mining of EHR data for research, the terms 'hacking', 'manipulation', and 'wrangling' should be outlawed"* Owen Johnson, University of Leeds

The framework[3] helps systematically identify many potential data quality issues and mark-up every data point affected. This facilitates detailed assessment of the data quality issues relevant to mining care-pathways.

Our structured approach saves time and brings rigor to the management and mitigation of DQ issues in process-mining research. The framework adds data quality information metadata to the research data and excludes specific data from particular experiments based on this metadata.

The resulting metadata is then used within cohort selection, experiments and process mining software so that our research with this data is based on data of known quality.



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)

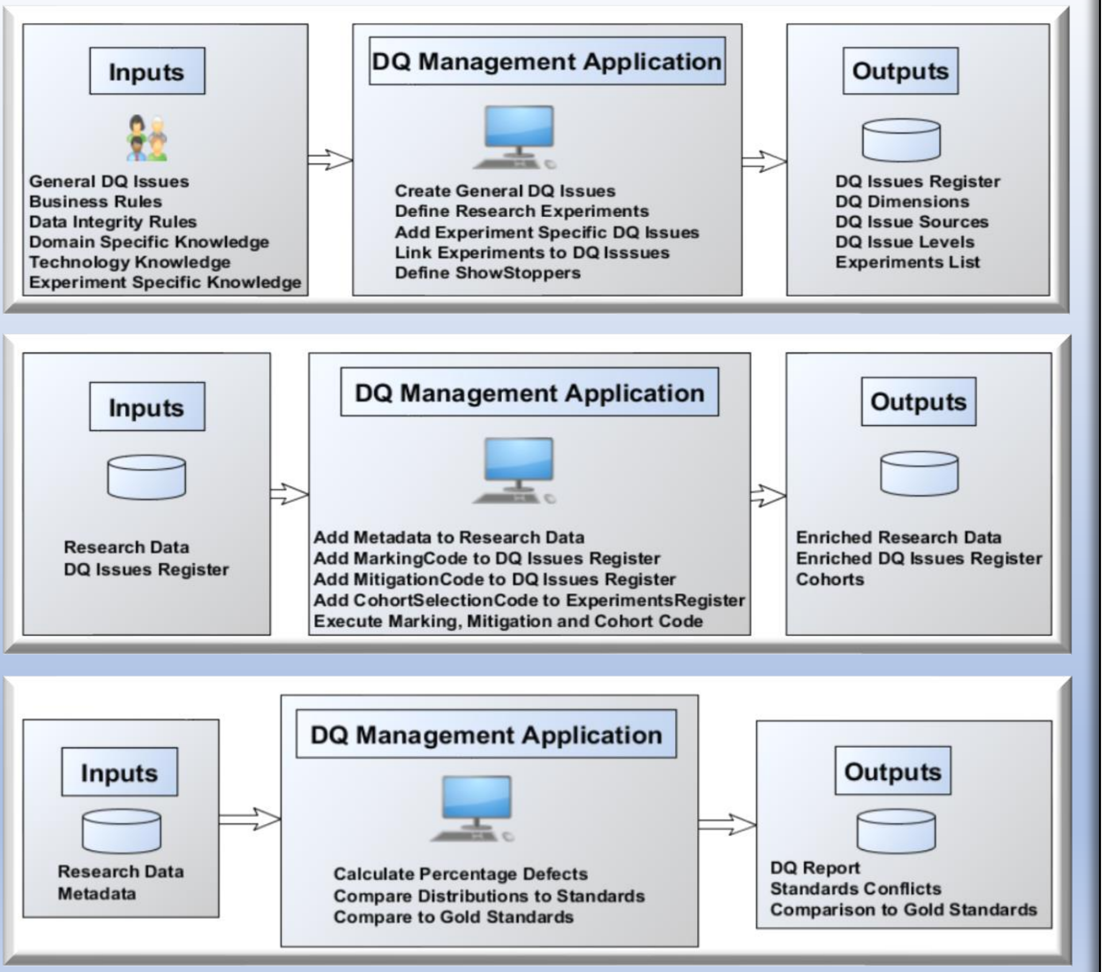
Figure 1: Event log with metadata added [3]

Using The Framework:

Phase 1:
Identify potential DQ Issues

Phase 2:
Apply the framework to the research data

Phase 3:
Report on Phases 1 & 2



Inputs

General DQ Issues
Business Rules
Data Integrity Rules
Domain Specific Knowledge
Technology Knowledge
Experiment Specific Knowledge

DQ Management Application

Create General DQ Issues
Define Research Experiments
Add Experiment Specific DQ Issues
Link Experiments to DQ Isssues
Define ShowStoppers

Outputs

DQ Issues Register
DQ Dimensions
DQ Issue Sources
DQ Issue Levels
Experiments List

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

Inputs

Research Data
Metadata

DQ Management Application

Calculate Percentage Defects
Compare Distributions to Standards
Compare to Gold Standards

Outputs

DQ Report
Standards Conflicts
Comparison to Gold Standards

Our Case Study: Applying the framework to dental EHR process mining research

The Data:

Extract from a dental EHR database from the Health Service Executive (South), Ireland, covering 41 clinics in two counties, Cork & Kerry. Research data extract covered public health school dental screenings.

School children (n = 231,760)
Clinical charts (n = 1,016,197)
Treatment events (n = 3,169,864)
Tooth conditions (n = 32,291,681)

Where are the DQ information sources?[3]

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

What are the dimensions of DQ?

Four broad DQ issues that exist in process mining event logs were identified: Missing Data, Incorrect Data, Imprecise Data, Irrelevant Data. [1][2][3]

What did we find?

We found and logged over 100 DQ issues. We assessed and marked the data for each of them, excluded the compromised data from our experiments if appropriate and reported on these steps.

Example DQ issues report:

DQ Issue Name (Integrity Rules)	Rows marked	%
Entries in PMTreatments must have Client	48330	1.52
Entries in PMChart must have Client	3267	0.32
PMProcedure GroupNames table entries	18316	0.57
CompletionDate > =1990-01-01 00:19:02.000	197352	6.22

Other DQ metrics used:

- Distributions
- Comparison to gold standard