DQUEEN: A Data Quality Assessment and managing tool for OMOP Common Data Model

Junghyun Byun, BE[1], SongHeui OH, MS[1], DongSu Park, BE[1],
Ji Young Hwang, EMT-P,Ph.D[1], Rae Woong Park, MD, Ph.D[1,2]

[1]Department of Biomedical Informatics, School of Medicine, Ajou University, Suwon, Korea;
[2]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea;

Is this the first time you have submitted your work to be displayed at any OHDSI Symposium?

Yes __o__ No._____

Abstract

Data quality is important for generation of accurate evidence from large amount of data, particularly in the Common Data Model (CDM), which is transformed from the source data of multiple institutions. We integrated the DQ checks from Achilles, PEDSnet, DQe-c and In-house DQ check. Total 1,255 checks were analyzed and 130 duplicate checks were eliminated. These checks were applied to the source data and CDM data for DQA. Based on this, we designed and developed the DQ tool for evaluating and managing DQ. By proving the visualizing information and score about DQ, This tool expected to be utilized the objective indicator.

Introduction

Each distributed research network (DRN) makes considerable effort to confirm the data quality (DQ) by providing their own DQ tools to ensure that the CDM data is "high-quality" or "ready for research use"[1]. However, although not significantly different, the existing DQ tools have different terminology, validation scope, and DQ checks and criteria. Furthermore, for most DQ tools, the source data are not regarded well; therefore, they suffer from limitations for confirming the quality of source data. The purpose of this study is to confirm the generalization possibility of DQA by applying the existing DQ checks and In-house DQ checks to the source data and OMOP-CDM. Moreover, we propose a DQ tool that can perform DQ evaluation between the source data and OMOP-CDM by providing DQA for the source data.

Method

Data quality concept and check review

We conducted a literature review about DQA on EHR data or OMOP-CDM for confirming the DQ concept and verified DQ checks of three DQ tools: Achilles[2], PEDSnet[3], DQe-c[4]. Each DQ tool have 185 checks from Achilles, 765 checks from PEDSnet, and 12 checks from DQe-c. Additionally, 704 In-house DQ checks were analyzed. After elimination of the duplicate DQ checks, we categorized them together using Kahn et al's[5] DQ concepts. Based on these DQ concepts, we applied these DQ checks unified to the source data and the OMOP-CDM data.

Data quality assessment process of DQUEEN

The system process is designed from the low to high complexity DQ concept, which identifies data errors. The flow step is constructed by 7 DQ concepts (Figure 1). First, verify Data uniqueness and distribution of missing data (Completeness & Uniqueness). Second, identify the conformance of data such as pre-defined format (Conformance). Third, check the accuracy of the data (Accuracy). Finally, evaluate the error about

time variable and the others like logical error (Plausibility). The result of DQA is provided in the form of numeric values, such as percentage or absolute values. Based on this result, the score information by DQ concepts or the visualizing information are provided. Based on this process, the system is designed by modularizing it such that it can be applied to the source data and the OMOP-CDM data.

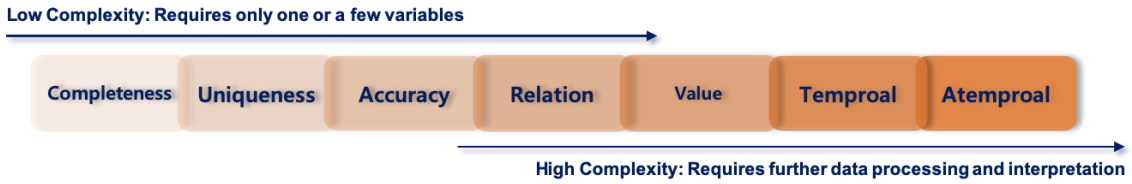

Figure 1. DQUEEN's Data Quality Assessment Flow.

Result

Integration of data quality concept and check

The total number of DQ checks is 1,255 (742 checks by the three DQ tools and 513 In-house checks). 321 duplicate DQ checks were excluded. We adopt 'Accuracy' for additional DQ concept. Table 1 shows the DQ checks applied to each field for DQA.

Table 1. Integrated Data Quality Concept and check in OHDSI, PEDSnet and DQe-c

DQ Concept	Subcategory	DQ checks				DQ checks for	
		Achilles	PEDSnet	DQe-c	In-house	Source data	CDM data
Plausibility	Temporal	1	72	-	31	51	48
	Atemporal	71	180	2	231	280	186
	Uniqueness	-	-	-	71	49	21
Completeness	Completeness	15	209	7	155	36	197
Conformance	Value	90	203	1	122	151	121
	Relation	-	19	2	43	53	11
Accuracy	Accuracy	-	-	-	51	35	16
Total		185	683	12	704	655	600

DQA Process of DQUEEN

The total process of DQUEEN is structured by the Meta module and CDM module (Appendix A).Based on the DQA results performed by each module, the score for each DQ concept for the table and schema was calculated. By visualizing the DQ score and evaluation results, the source data and the CDM data can be compared.

Conclusion

This study is determined the generalization possibility of DQA by applying existing DQ checks and In-house DQ checks to the source and CDM data. Moreover, by providing DQ evaluation using score and visualization, this tool was expected to be used as an objective quality index. However, as this tool cannot be used to evaluate the quality of unstructured data, further studies are needed to measure the quality of unstructured data.

Acknowledgement

This work was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [grant number : HI16C0992] and the Bio Industrial Strategic Technology Development Program (20003883) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

1. Kahn, M. G., Raebel, M. A., Glanz, J. M., Riedlinger, K., & Steiner, J. F. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Medical care, 2012;06:50(0).
2. Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) - descriptive statistics about a OMOP CDM database [Internet] Available from: https://github.com/OHDSI/Achilles.
3. The PEDSnet Data Quality Assessment Toolkit (OMOP CDM) [Internet] Available from: https://github.com/PEDSnet/Data-Quality-Analysis.
4. DQe-c [Internet] Available from: https://github.com/hestiri/DQe-c.
5. Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., ... & Liaw, S. T. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. Egems, 2016;09:4(1).

Appendix A. The total process of DQUEEN