Data Quality Dashboard v2.0

Clair Blacketer[1,2], Frank DeFalco[1], Anthony Molinaro[1], Dmitry Ilyn[3], Luis Alaniz[3], Maxim Moinat[2,4]

1. Janssen Pharmaceutical Research and Development LLC, Titusville, NJ, USA
2. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, NL
3. Odysseus Data Services, Prague, Czech Republic
4. The Hyve, Utrecht, NL

Background

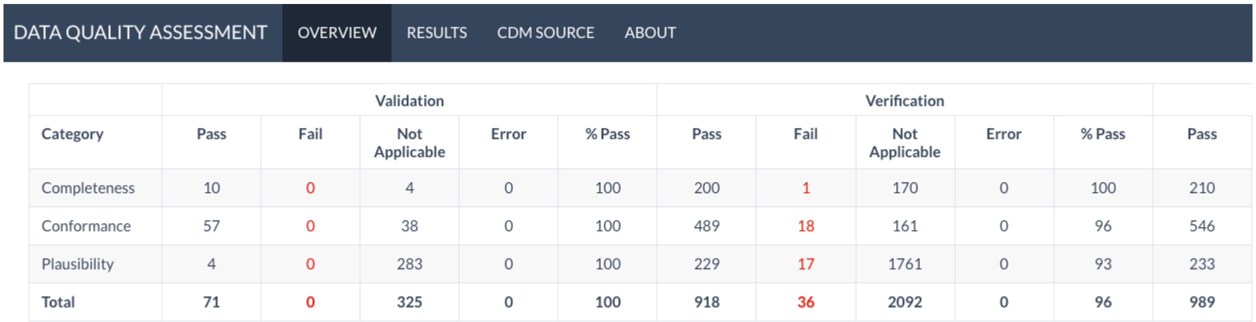
The Data Quality Dashboard (DQD) was presented to the Observational Health Data Sciences and Informatics (OHDSI) community in 2019 at the global symposium (1). Since then, it has become a mainstay for Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) experts and observational health data networks, both federated and centralized. The European Health Data and Evidence Network (EHDEN) relies on it heavily to evaluate the myriad of databases around Europe that contribute to the network (2). The National COVID Cohort Collaborative (N3C) uses it as a basis for the quality measures applied to the data they receive and network studies conducted across the OHDSI community are starting to require DQD results shipped alongside study results (3). This level of adoption is vital to the life and enhancement of an application as new users reveal the limitations of the current state and offer fresh ideas to improve the usability of the tool. DQD v2.0 includes an updated method for summarizing the results of the data quality checks, and a new shiny application that presents the results in a more meaningful way.

Methods

We collected feedback over the course of two years from members of the global community who regularly use the DQD, and two common themes emerged. First, data quality checks, as they are evaluated in the results, are given an outcome of either “Pass” or “Fail”. This simple, binary classification was deemed to be problematic. The very nature of a Common Data Model, or a model used to standardize similar types of data, is that provisions need to be made for many types of data that each user may or may not have in their own database. For example, the OMOP CDM contains both a Note and Note NLP meant to house unstructured health data derived from free text fields. Few community members have access to this type of data and fewer still choose to standardize it. When faced with empty tables the DQD gives an unequivocal “Pass” to all quality checks applied. This then inflates the final total of the number of quality checks passed by a given database, lulling the user into a false sense of data quality prowess. In v2.0, however, we account for this grade inflation by introducing two new check outcomes. Instead of just “Pass” or “Fail”, there is now the option for quality checks to be labelled as “Pass”, “Fail”, “Error”, or “Not Applicable” (figure 1). “Error” indicates that there was an error somewhere in a SQL statement and “Not Applicable” indicates that there was no data to supply a given quality check.

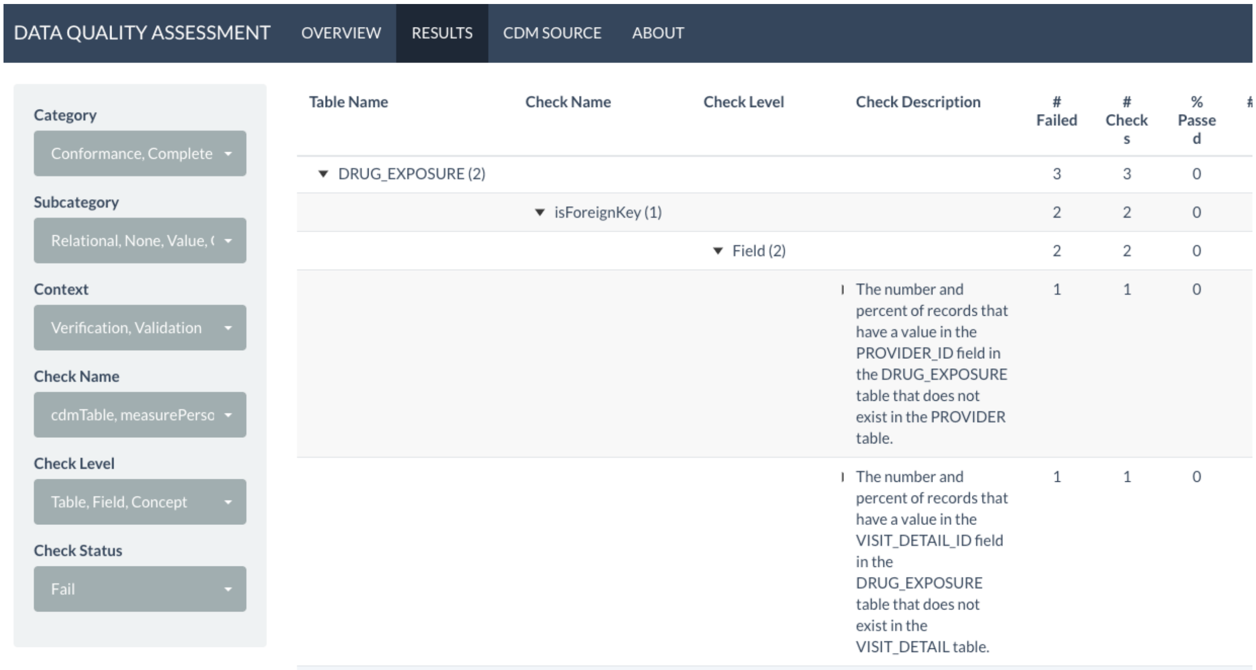
The second theme we heard often was related to the web application designed to display the results. It was originally written in javascript and html wearing an RShiny trenchcoat, meaning it was not taking full advantage of either platform. The biggest issue here was the inability to group data quality issues by table. We observed that most data owners prefer to walk through data quality failures table-by-table, checking each off before moving to the next. The DQD v2.0 facilitates this desire by providing a new interface fully written in RShiny that allows the user to group the data quality checks in multiple ways, giving a better sense of the overall health of the database and where to focus cleaning efforts (figure 2).

Results



	Validation					Verification					
Category	Pass	Fail	Not Applicable	Error	% Pass	Pass	Fail	Not Applicable	Error	% Pass	Pass
Completeness	10	0	4	0	100	200	1	170	0	100	210
Conformance	57	0	38	0	100	489	18	161	0	96	546
Plausibility	4	0	283	0	100	229	17	1761	0	93	233
Total	**71**	**0**	**325**	**0**	**100**	**918**	**36**	**2092**	**0**	**96**	**989**

Figure 2. A summary of DQD results showing the columns for "Error" and "Not Applicable".



Category: Conformance, Complete

Subcategory: Relational, None, Value, (

Context: Verification, Validation

Check Name: cdmTable, measurePerso

Check Level: Table, Field, Concept

Check Status: Fail

Table Name	Check Name	Check Level	Check Description	# Failed	# Checks	% Passed	#
▼ DRUG_EXPOSURE (2)				3	3	0	
	▼ isForeignKey (1)			2	2	0	
		▼ Field (2)		2	2	0	
			The number and percent of records that have a value in the PROVIDER_ID field in the DRUG_EXPOSURE table that does not exist in the PROVIDER table.	1	1	0	
			The number and percent of records that have a value in the VISIT_DETAIL_ID field in the DRUG_EXPOSURE table that does not exist in the VISIT_DETAIL table.	1	1	0	

Figure 2. DQD results by CDM table.

Conclusion

Version 2 of the Data Quality Dashboard provides updates to the application that make it easier to interpret by introducing two new check outcomes. The newly added RShiny application allows users more control over how the results are displayed, organizing them in a manner that tells you the overall health of the database at a glance.

References/Citations

1. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. Journal of the American Medical Informatics Association : JAMIA. 2021 Jul 27;

2. Blacketer C, Voss EA, DeFalco F, Hughes N, Schuemie MJ, Moinat M, et al. Using the Data Quality Dashboard to Improve the EHDEN Network. Applied Sciences. 2021 Jan;11(24):11920.

3. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021 Mar 1;28(3):427–43.