Assessing data quality at DARWIN EU® data partner onboarding

Sofia Bazakou[2], Anne van Winzum[2], Maxim Moinat[1]
[1]Erasmus Medical Center, Rotterdam, The Netherlands
[2]The Hyve, Utrecht, The Netherlands

Background

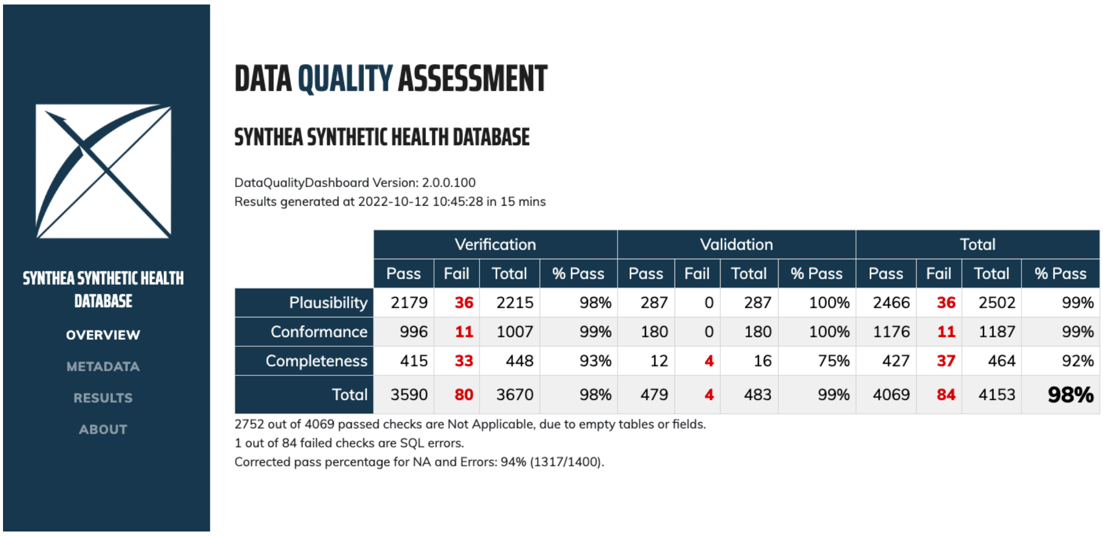
DARWIN EU® (Data Analysis and Real-World Interrogation Network) is setting up a network of health care databases across Europe to support large scale regulatory studies. These databases have been standardised to the OMOP CDM and need to be assessed for conformance and quality before studies can be executed. This assessment is carried out by the DARWIN EU® Network Operations pillar and is aiming to inform the European Medicines Agency (EMA) and the DARWIN EU® Study Operations pillar about the strengths and weaknesses of the network.

Data quality is strongly context dependent therefore data quality assessments should be performed at different stages in the evidence generation workflow. During onboarding of a data source in the DARWIN EU® data network the DARWIN EU® Network Operations pillar collects metadata from the data partners in a proactive manner, using a combination of OHDSI and bespoke tools. This process is to be repeated with each data release.

Methods

The data partners selected to be onboarded are required to a run a Data Quality Assurance Package directly against the OMOP CDM data to produce aggregated information that can be shared with the DARWIN EU® Coordination Centre (CC). This package consists of three tools (R packages) – the OHDSI Data Quality Dashboard (DQD), DARWIN EU CDMOnboarding and DARWIN EU DashboardExport. The latter two depend on the results from the OHDSI Achilles package [1].

1. The Data Quality Dashboard was developed and is maintained by the OHDSI community. This tool uses a systematic approach to evaluate over 4,000 data quality checks against a given CDM instance [2] and it has proven to be an invaluable tool in EHDEN during the mapping process of data sources. The results helped identifying multiple issues that were then subsequently addressed and it provided opportunities for training on data standardisations [3]. The results are saved as a machine-readable JSON file that can be shared by the DARWIN EU® data partners.



DATA QUALITY ASSESSMENT

SYNTHEA SYNTHETIC HEALTH DATABASE

DataQualityDashboard Version: 2.0.0.100
Results generated at 2022-10-12 10:45:28 in 15 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	2179	36	2215	98%	287	0	287	100%	2466	36	2502	99%
Conformance	996	11	1007	99%	180	0	180	100%	1176	11	1187	99%
Completeness	415	33	448	93%	12	4	16	75%	427	37	464	92%
Total	3590	80	3670	98%	479	4	483	99%	4069	84	4153	98%

2752 out of 4069 passed checks are Not Applicable, due to empty tables or fields.
1 out of 84 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 94% (1317/1400).

Figure 1: Main component of the Data Quality Dashboard

2. The CdmOnboarding package [4], developed by DARWIN EU®, provides insight into the completeness, transparency and quality of the performed Extraction Transform, and Load (ETL) process and the readiness of the data partner to be onboarded in the DARWIN EU® data network and the participation in research studies [Rijnbeek, Moinat, Introduction, para. 2]. The CdmOnboarding package generates a report as a Word document directly from the OMOP CDM and uses results generated by Achilles. The quality checks applied by the CdmOnboarding package consist of two parts: one focused on clinical data and one on the vocabulary data and provides an additional quality control check on top of the DQD checks. The package is based on the CDMInspection package developed by the EHDEN consortium[5].

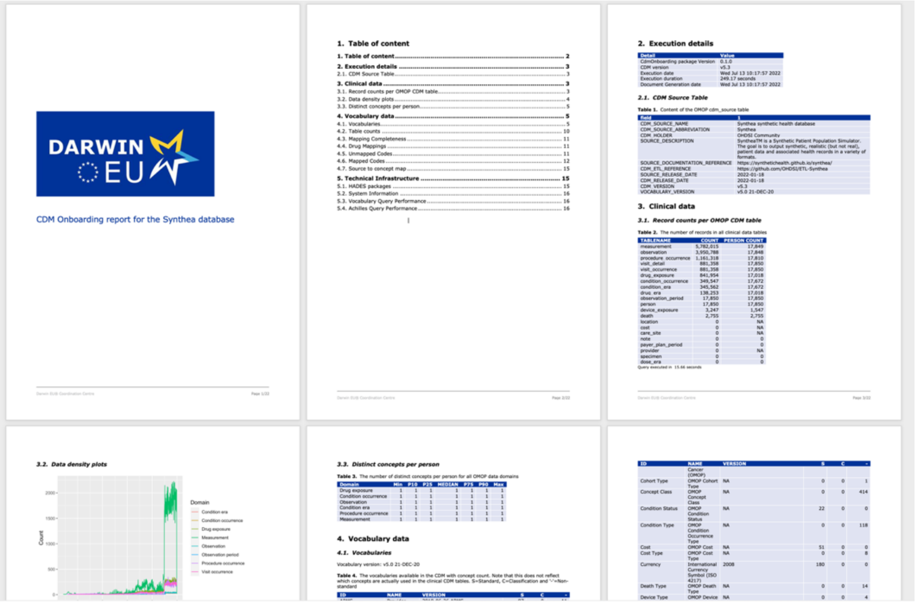
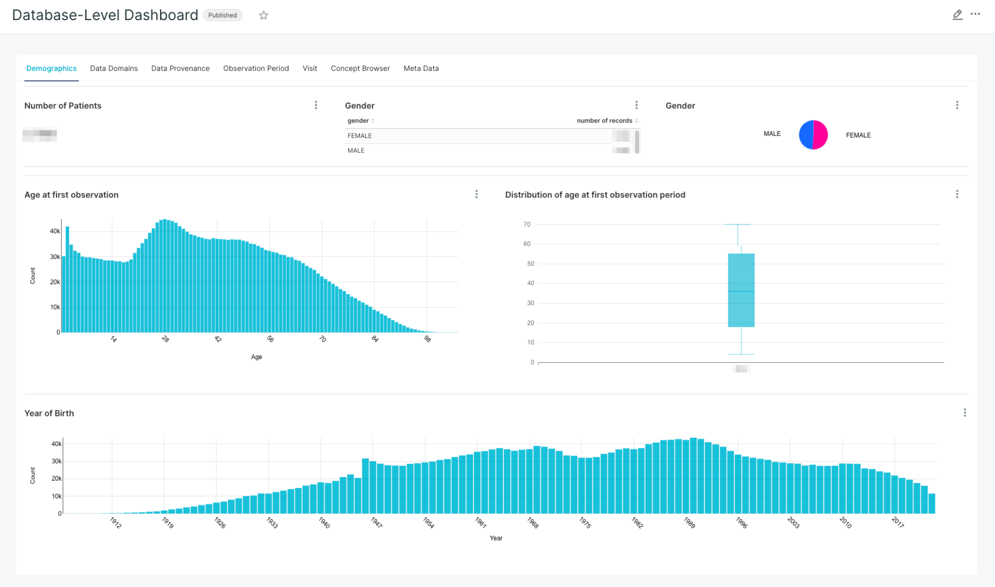


Figure 2: The first pages of an example CdmOnboarding report.

3. The DashboardExport package [6], developed by DARWIN EU®, exports a subset of Achilles analysis results and rounds results to the nearest 100. These results are imported into the DARWIN EU® Portal, to be visualised in the Database Dashboard.



Number of Patients: [redacted]

gender	number of records
FEMALE	[redacted]
MALE	[redacted]

Figure 3: DARWIN EU® Database Dashboard, populated with results from the DashboardExport package

The Data Quality Assurance Process applied during the onboarding of data sources consists of multiple steps. Firstly, the data partner is required to fill out an Onboarding Document with questions on the data quality assurance processes being part of this document. The data partner is asked to give insights into their quality control process, both on their source data and on the data transformed to the OMOP CDM. The data partner is also required to run the quality control packages described above. The results of the quality control packages are shared by the data partner together with the Onboarding document. The DARWIN EU® CC reviews the submitted documents, which informs the recommendation on the acceptance of a data partner into the DARWIN EU® network. The CC provides feedback to the data partner and tracks the cause, possible remediations and status of each data quality issue in a dedicated tracker.

After the initial data quality assurance (QA) process, the QA process is routinely repeated. The packages are to be rerun by the data partners after every CDM update, triggered by additional source data ingestion, ETL script changes, new version of the OMOP vocabularies or OMOP CDM.

Results

In December 2022, we completed the data quality process for the first ten data partners in the DARWIN EU® network. We evaluated the QA documents and were able to identify data quality issues for each data partner, which were given as feedback together with suggested remediations. Since the databases vary widely (type, country, collection process, etc.) so are the issues found. Yet there are some outcomes observed throughout the network.

Conclusion

The process described above is applied as part of the onboarding process for each DARWIN EU® data partner. The results of the QA process were part of the recommendation for inclusion of the data partner in the DARWIN EU® network and are very beneficial. From the data partners' position, their OMOP CDM instance can be improved in further iterations and in addition they can be aware of more universal challenges, creating a strong network. And from the CC perspective, after the end of the onboarding process, it allows for quicker and more reliable evaluation of databases for study participation and a feedback mechanism for the quality assessment to be improved with every new onboarding and new study.

The Quality Assurance Package is constantly updated, e.g., to cover new quality checks that may become of value when the Catalogue of Analytics grows and recommendations coming from the data quality framework delivered by EMA. One example of the former is the incorporation of drug exposure diagnostics for selected ingredients. The framework we developed can be extended to meet our future goals.

References

1. https://github.com/ohdsi/achilles
2. Clair Blacketer, Frank J Defalco, Patrick B Ryan, Peter R Rijnbeek, Increasing trust in real-world evidence through evaluation of observational data quality, *Journal of the American Medical Informatics Association*, Volume 28, Issue 10, October 2021, Pages 2251–2257, https://doi.org/10.1093/jamia/ocab132
3. Blacketer, C.; Voss, E.A.; DeFalco, F.; Hughes, N.; Schuemie, M.J.; Moinat, M.; Rijnbeek, P.R. Using the Data Quality Dashboard to Improve the EHDEN Network. *Appl. Sci.* **2021**, *11*, 11920. https://doi.org/10.3390/app112411920
4. https://github.com/darwin-eu/CdmOnboarding
5. https://github.com/EHDEN/cdminspection
6. https://github.com/darwin-eu/DashboardExport