Brazilian administrative data for real-world research: a deterministic linkage procedure and OMOP CDM harmonization

Jessica Mayumi Maruyama[1], Julio Cesar Barbour Oliveira[1]
[1] Precision Data

Background

Real-world evidence (RWE) refers to evidence derived from real-world data (RWD) and settings, including routine clinical practice, electronic health records (EHRs), patient registries, and other sources of data outside of traditional clinical trials.[1-3] RWE provides valuable insights into the safety, effectiveness, and value of healthcare interventions in real-world patient populations and allows the examination of patient journeys in real-world practice.[1-3] Like many countries, Brazil has a large volume of health data collected via national information systems and available in different databases.[3-5] The Brazilian National System, known as SUS (*Sistema Único de Saúde*) provides health care to approximately 200 million inhabitants, of which 75% depend exclusively on SUS.[3-5] The national administrative database is referred to as DATASUS and is publicly available through the Brazilian Ministry of Health website.[4,6] Information collected by DATASUS includes statistics from all municipalities across the country encompassing multiple data sources systems, such as hospital and ambulatory data.[4] However, the multiple systems within DATASUS are not integrated and data collected by different health services lack unique key identifiers at the individual level, which represents a challenge for the record linkage of such datasets.[4,5] Moreover, to the best of our knowledge, there is only one previous research evaluating the harmonization of Brazilian claims data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), which was focused on COVID-19 records.[6] Therefore, the current study aimed to describe the methods and the partial results of the minimal viable product (MVP) of parameter setting needed to create a dataset using Brazilian claims data, evaluating the data quality against an OMOP CDM instance using the DataQualityDashboard.[7]

Methods

Description of databases: The DATASUS databases used were the Hospital Information System ([*Sistema de Informação Hospitatalar]* – SIH) and the Ambulatory Information System ([*Sistema de Informação Ambulatorial]* – SIA) from January 2008 to present date. We are currently developing a data pipeline, which will undergo monthly updates. It is important to note that while these data are updated every month, there is a delay of two months in the publication of reports. This means that the data available today (June) will include information up until April. SIH captures all inpatient personal information, procedures, treatments and separation (hospital discharges, transfers, and deaths). Examples of included variables are sex, age, number of hospitalizations, ICD-10 codes of primary and secondary diagnosis, residential address, zip code, length of stay, and total amount of reimbursed hospital services. SIA is divided into two subsystems: the individual outpatient production report (BPAI), and the authorization of high-complexity outpatient procedures (APAC). The information available in both datasets includes all outpatient procedures, consultations, ICD-10 codes of a primary and secondary diagnosis, medicines dispensed for domiciliary use, and personal data, such as date of birth, residential address, zip code, and sex. More information on SIH and SIA was published elsewhere.[4,5] All publicly available individual-level data in DATASUS are anonymized and encrypted.[8]

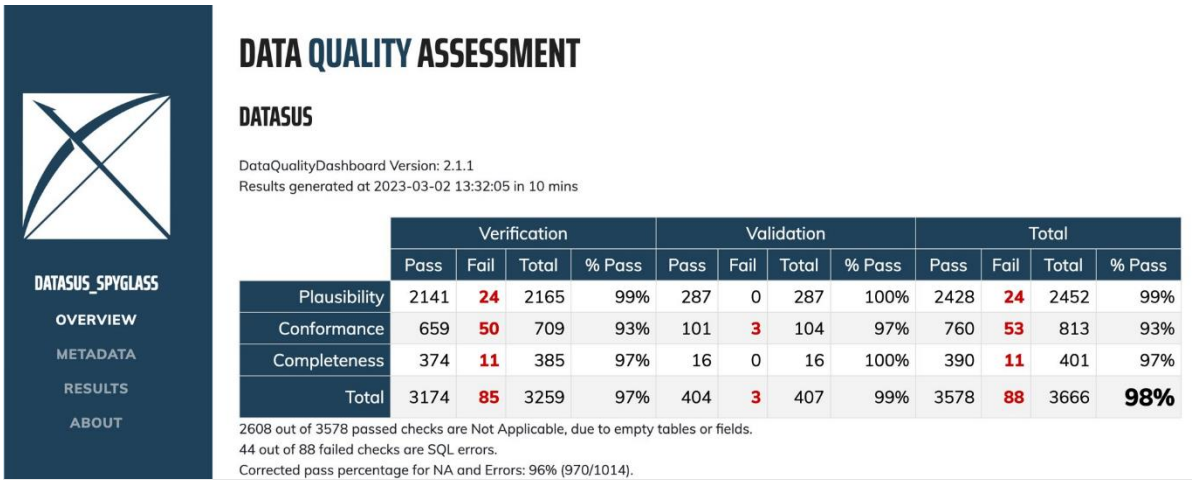
Record linkage and OMOP CDM harmonization: All original records from the DATASUS database were

analyzed to assess the consistency of information for a unique patient key. Patients with inconsistencies in basic information (e.g., date of birth or gender) were excluded from the database. Patients with different primary keys but with matching basic information such as zip code, date of birth, and gender, were also excluded from the analysis because these data were used to create the composite key to proceed with database linkage. After the cleaning and pre-processing stage, a deterministic linkage algorithm was developed to connect hospitals with outpatient records using the key information of zip code, date of birth, and gender. During this process, patients from zip codes with more than 2500 distinct individuals linked to them in the database were also excluded. A total of 5.82 million patients were included in the final dataset. Following the establishment of this patient cohort and the subsequent linkage with the hospital database, a standardized dataset encompassing the complete health history of these 5.82 million patients was constructed. Subsequently, this dataset underwent a transformation into the OMOP CDM model. The process of standardization into the OMOP process adhered to the guidelines provided by the online community manuals. The data quality check was conducted using the DataQualityDashboard version 2.1.1.[7]

Results

We present a partial result of our OMOP MVP. It is important to clarify that the current results are based on a subset of the DATASUS database that has been mapped to OMOP, indicating the quality of the mapping process. While we are in the process of mapping the entire database, consisting of 5.82 million patients, the results we are presenting here are derived from a sample of this patient population. We intend to have a complete database result to showcase in the form of a poster presentation by the time of the symposium.

Figure 1 shows the Data Quality Dashboard of our MVP OMOP DATASUS. Out of these 2,165 checks evaluating data plausibility, only 24 failed in our database, while 2,141 passed. Additionally, there are 3,666 queries assessing the quality mapping of the OMOP database. Among these, we encountered 88 failures and successfully passed 3,578 queries. Our database showed an overall pass rate of 98%, indicating a satisfactory index of the mapping quality. However, it is important to acknowledge that not all tables (e.g., visit_detail, specimen, and note_nlp) have been mapped to the OMOP model, which introduces a confirmation bias in this index.



	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	2141	24	2165	99%	287	0	287	100%	2428	24	2452	99%
Conformance	659	50	709	93%	101	3	104	97%	760	53	813	93%
Completeness	374	11	385	97%	16	0	16	100%	390	11	401	97%
Total	3174	85	3259	97%	404	3	407	99%	3578	88	3666	**98%**

2608 out of 3578 passed checks are Not Applicable, due to empty tables or fields.
44 out of 88 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 96% (970/1014).

Figure 1. Data Quality Dashboard of our MVP OMOP DATASUS.

Conclusion

The results of the current study demonstrated a data treatment methodology for DATASUS datasets and the provision of a new database that facilitates collaboration within the OHDSI community. Further initial studies utilizing real-world data from this dataset, such as a descriptive study investigating patient outcomes associated with the use of reprocessed duodenoscopes for endoscopic retrograde cholangiopancreatography (ERCP) procedure, have already been conducted by the authors. This study was also submitted as a poster presentation at the Symposium. In the future, we anticipate the emergence of more studies and insights derived from this database, which will significantly enhance our understanding of various healthcare aspects.

References

1. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022;22:287 https://doi.org/10.1186/s12874-022-01768-6.
2. Nabhan C, Klink A, Prasad V. Real-world evidence—what does it really mean? JAMA Oncol. 2019;5(6):781.
3. Justo N, Espinoza MA, Ratto B, et al. Real-world evidence in healthcare decision making: global trends and case studies from latin america. Value Health. 2019;22(6):739-749.
4. Ali MS, Ichihara MY, Lopes LC, Barbosa GCG, Pita R, Carreiro RP, Dos Santos DB, Ramos D, Bispo N, Raynal F, Canuto V, de Araujo Almeida B, Fiaccone RL, Barreto ME, Smeeth L, Barreto ML. Administrative data linkage in Brazil: potentials for health technology assessment. Front Pharmacol. 2019;10:984. doi: 10.3389/fphar.2019.00984.
5. Junior AAG, Pereira RG, Gurgel EI, Cherchiglia M, Dias LV, Ávila J, et al. Building the national database of health centred on the individual: administrative and epidemiological record linkage - Brazil, 2000-2015. Int J Popul Data Sci. 2018;3(1). doi: 10.23889/ijpds.v3i1.446.
6. Junior EPP, Normando P, Flores-Ortiz R, et al. Integrating real-world data from Brazil and Pakistan into the OMOP common data model and standardized health analytics framework to characterize COVID-19 in the Global South. J Am Med Inform Assoc. 2023;30(4):643-655. doi:10.1093/jamia/ocac180
7. DATASUS – Ministério da Saúde [Internet]. [Brasília] Ministério da Saúde (BR). [cited 2023 May 30]. Available from: https://datasus.saude.gov.br/
8. Blacketer C, Schuemie FJ, Ryan PB, Rijnbeek P.“Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc 2021;28(10):2251-2257. https://doi.org/10.1093/jamia/ocab132.