Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220430

Making EHRs Trustable: A Quality Analysis of EHR-Derived Datasets for COVID-19 Research

Miguel PEDRERA-JIMENEZ[a,b,1], Noelia GARCIA-BARRIO[a], Paula RUBIO-MAYO[a], Guillermo MAESTRO[c], Antonio LALUEZA[c], Ana GARCIA-REYNE[c], María José ZAMORRO[c], Alejandra PONS[c], María Jesús SANCHEZ-MARTIN[c], Jaime CRUZ-ROJO[a], Víctor QUIROS[a], José María AGUADO[c], Juan Luis CRUZ-BERMUDEZ[a], José Luis BERNAL[a], Laura MERSON[d], Carlos LUMBRERAS[c] and Pablo SERRANO[a]
[a]*Data Science Unit, Research Institute Hospital 12 de Octubre, Madrid, Spain.*
[b]*ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain.*
[c]*Internal Medicine Department, Hospital Universitario 12 de Octubre, Madrid, Spain.*
[d]*ISARIC Global Support Centre, University of Oxford, Oxford, UK.*

Abstract. One approach to verifying the quality of research data obtained from EHRs is auditing how complete and correct the data are in comparison with those collected by manual and controlled methods. This study analyzed data quality of an EHR-derived dataset for COVID-19 research, obtained during the pandemic at Hospital Universitario 12 de Octubre. Data were extracted from EHRs and a manually collected research database, and then transformed into the ISARIC-WHO COVID-19 CRF model. Subsequently, a data analysis was performed, comparing both sources through this convergence model. More concepts and records were obtained from EHRs, and PPV (95% CI) was above 85% in most sections. In future studies, a more detailed analysis of data quality will be carried out.

Keywords. Electronic Health Records, Real World Data, Data Quality, Completeness, Correctness, Semantics, Standards, ISARIC-WHO, COVID-19.

1. Introduction

Electronic health records (EHRs) are conceived as a digital repository of health data that is used for individual patient healthcare [1]. In addition, it can be applied to other purposes, known as secondary uses, including clinical research and public health [2]. To achieve this, recent studies have designed methodologies based on health information standards for allowing the effective reuse of EHRs, which is essential for obtaining data in an agile, flexible and efficient way [3, 4]. Traditionally, data for research have been manually collected in purpose-built and controlled databases. This made it is necessary to audit the quality of data obtained from EHRs through systematic and validated methodologies [5, 6]. This became evident during the COVID-19 pandemic when two studies, published in high-impact journals, had to be retracted due to data quality, among other issues (10.1016/S0140-6736(20)31180-6 and 10.1056/NEJMoa2007621).

[1] Corresponding Author, Miguel Pedrera Jiménez, Data Science Group, Hospital Universitario 12 de Octubre, Av. de Andalucía S/N, 28041, Madrid, España; E-mail: miguel.pedrera@salud.madrid.org.

Thus, this study analyzes data quality of an EHR-derived dataset for COVID-19 research, obtained at Hospital Universitario 12 de Octubre, Madrid, Spain (H12O) [7].

2. Methods

The study compares two study databases: one obtained only from structured EHRs, and other manually transcribed from structured EHRs, clinical reports, and external sources. Data for both databases were collected between March 2020 and September 2020. This work is part of the research line on reuse of EHRs at H12O [3, 4, 7-9].

2.1. Extraction, transformation and loading of health data

The first source for data extraction was the structured EHRs of H12O, which have been modeled and formalized through health information standards including ISO 13606 [10], and controlled terminologies such as SNOMED CT and LOINC [11, 12]. This effort has allowed the full meaning reuse of EHRs, without additional manual efforts, in data collection processes for research [3]. Hence, it was possible to participate in different data-driven projects during the COVID-19 pandemic, including TriNetX, EHDEN Consortium, 4CE Consortium and ISARIC Consortium [4, 7].

The second data source was the data collected within the STOP-CORONAVIRUS project, a clinical study on COVID-19 developed at H12O [13]. In this study, a specific data collection was carried out in a relational database implemented in REDCap [14], being manually transcribed from structured EHRs, clinical reports and external sources.

Data from both sources were transformed into a common model for analysis and comparison. The COVID-19 case report form (CRF) proposed by ISARIC-WHO was chosen as convergence model, due to its international adoption for COVID-19 clinical data collection, and because H12O participates in the ISARIC Consortium by transferring EHRs without manual efforts. [15, 16]. Thus, two identical databases based on this model were implemented in REDCap [14], and then the data from the same cohort of 1732 COVID-19 patients were loaded into them from each data source.

2.2. Data quality analysis of health data

Determining the quality of health data is not a straightforward task since it can be measured from different perspectives. In the review carried out by Weiskopf et al., five data quality dimensions were identified, as well as seven methods to evaluate them [5]. Based on this review, the 'gold-standard comparison' method was selected for the analysis, establishing the STOP-CORONAVIRUS dataset as the reference against which to compare data obtained from EHRs. Two data quality dimensions were analyzed:

- **Completeness**, i.e., if a fact about a patient is recorded in EHRs. For this, both data sources were analyzed to determine the coverage of the concepts specified by ISARIC-WHO, the volume of data obtained (patient-level records), and the cohort coverage achieved.
- **Correctness**, i.e., if a record present in EHRs is true. For this, both data sources were compared to determine, for each equivalent (patient, date of registration and concept) non-null record, whether EHRs report same content as the gold-standard. Thus, the Positive Predictive Value (PPV, 95% CI) was calculated.

Hence, an algorithm was implemented with R programming language [17], with which it was determined, for each patient and record, if the data exists in both sources, and if so, whether they match (same data type and content). Figure 1 shows the flowchart of the algorithm.

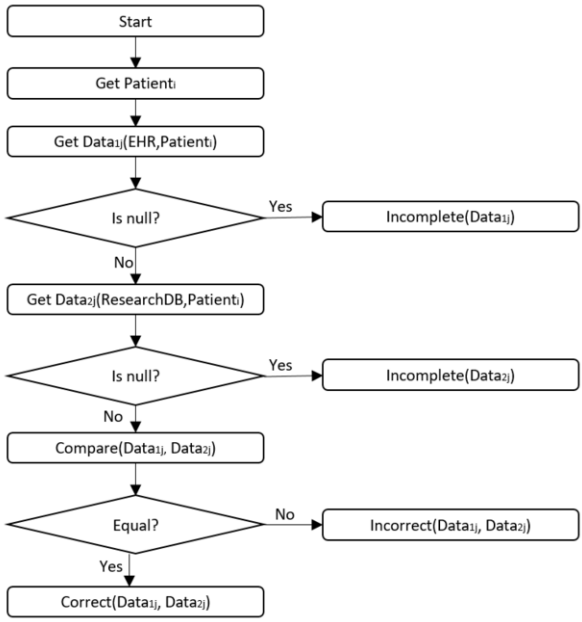

Figure 1. Algorithm for data quality analysis for completeness and correctness dimensions.

3. Results

3.1. Completeness analysis

The first result obtained was the completeness analysis of EHRs vs. the research database. Table 1 shows, grouped by section of the ISARIC-WHO CRF, the percentage of the CRF concepts and cohort covered, and the absolute volume of records obtained.

Table 1. Completeness analysis results.

ISARIC CRF Section	EHRs			Research database		
	Concepts (%)	Cohort (%)	Records (N)	Concepts (%)	Cohort (%)	Records (N)
Demographics	42.11	99.88	10,877	26.32	100	8051
Onset & admission	100	99.88	5266	82.35	100	3420
Signs and symptoms	53.85	99.88	12,714	30.77	100	20,177
Pre-admission medication	58.33	99.88	12,110	33.33	100	5318
Co-morbidities	100	99.88	34,600	75	100	24,948
Treatment	80	99.88	17,886	64	100	15,307
Complications	100	99.88	50,170	88.89	100	13,727
Clinical diagnostics	53.33	99.88	10,474	53.33	100	9998
Microbiology diagnostics	66.67	99.88	13,644	66.67	92.44	8511
Medication	82.35	99.88	27,554	61.76	100	20,285
Daily observations	86.96	99.88	331,796	30.43	99.48	27,848
Outcome	100	99.88	5190	66.67	100	3377

3.2. Correctness analysis

The second result obtained was the analysis of the correctness of EHRs compared with those recorded manually in the research database. Table 2 shows, for each section of the ISARIC-WHO CRF, the PPV (95% CI), as well as the number of concepts and records that could be compared.

Table 2. Correctness analysis results.

ISARIC CRF Section	Concepts compared	Records compared	PPV (95% CI)
Demographics	5	7505	97.79 (97.46, 98.12)
Onset & admission	2	1793	91.02 (89.70, 92.34)
Signs and symptoms	9	5824	58.41 (57.14, 59.68)
Pre-admission medication	4	5316	91.99 (91.26, 92.72)
Co-morbidities	15	24,933	91.36 (91.01, 91.71)
Treatment	15	13,805	84.80 (84.20, 85.40)
Complications	8	13,715	88.05 (87.51, 88.59)
Clinical diagnostics	5	5384	84.77 (83.81, 85.73)
Microbiology diagnostics	4	5008	99.16 (98.91, 99.41)
Medication	19	19,336	73.29 (72.67, 73.91)
Daily observations	14	19,690	78.70 (78.13 ,79.27)
Outcome	2	3375	92.74 (91.86, 93.62)

4. Discussion

The completeness analysis showed that, although the cohort was mostly covered from both sources, more concepts and volume of records were obtained from EHRs, since they were recorded during and for patient's care, rather than in an additional effort based on a fixed design. This was most evident in the 'Daily observations' section, with 86.96% concepts and 331,796 records in EHRs, vs. 30.43% and 27,848 in the research database.

Likewise, the correctness analysis showed that the EHRs has a PPV (95% CI) over 85% in most sections. The sections 'Signs and Symptoms', 'Medication' and 'Daily observations' were between 58% and 80%, which, thanks to this analysis, could be identified as gaps in the standardization and coverage of the EHRs (mainly due to free-text data entry). This allowed improving these information domains and thus obtaining higher quality data in future processes of obtaining EHRs-derived datasets for research.

These results highlight the value of the EHRs as a useful and valid source for research, in a scenario where multiple projects propose automated upload processes from healthcare information systems to databases and repositories for research [18, 19].

5. Conclusions

In this study, a quality analysis of EHR-derived datasets for COVID-19 research was performed. To this end, data were extracted from two sources: EHRs of H12O and a manually-collected research database. Then, both datasets were transformed into the ISARIC-WHO COVID-19 CRF model for comparative analysis. Thus, it could be concluded that the EHRs are more complete than a specific research database, and these data collected during the healthcare activity have an adequate accuracy.

In future studies, expanded and more detailed analysis will be performed, including the results for each of the concepts of the ISARIC-WHO CRF model.

Acknowledgment

This work has been supported by Research Projects PI18/00981, PI18/01047 and COV20/00181; funded by Instituto de Salud Carlos III, co-funded by ERDF/ESF.

We want to thank the ISARIC Consortium for its innovative vision in accepting our proposal for automated data upload from EHRs. We also want to thank Mª Teresa García Morales and Agustín Gómez de la Cámara (Scientific Support Unit, "i+12" Institute).

References

[1] Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform*. 2008;77(5):291-304. doi:10.1016/j.ijmedinf.2007.09.001.

[2] Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc*. 2007;14(1):1-9. doi:10.1197/jamia.M2273

[3] Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, et al.Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. *J Biomed Inform*. 2021;115:103697. doi:10.1016/j.jbi.2021.103697.

[4] Pedrera M, Garcia N, Rubio P, Cruz JL, Bernal JL, Serrano P. Making EHRs Reusable: A Common Framework of Data Operations. *Stud Health Technol Inform*. 2021;287:129-133. doi:10.3233/SHTI210831.

[5] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20(1):144-151. doi:10.1136/amiajnl-2011-000681.

[6] Kohane IS, Aronow BJ, Avillach P, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. *J Med Internet Res*. 2021;23(3):e22219. Published 2021 Mar 2. doi:10.2196/22219.

[7] Pedrera M, Garcia N, Blanco A, et al. Use of EHRs in a Tertiary Hospital During COVID-19 Pandemic: A Multi-Purpose Approach Based on Standards. *Stud Health Technol Inform*. 2021;281:28-32. doi:10.3233/SHTI210114.

[8] Pedrera M, Serrano P, Terriza A, et al. Defining a Standardized Information Model for Multi-Source Representation of Breast Cancer Data. *Stud Health Technol Inform*. 2020;270:1243-1244. doi:10.3233/SHTI200383.

[9] González L, Pérez-Rey D, Alonso E, et al. Building an I2B2-Based Population Repository for Clinical Research. *Stud Health Technol Inform*. 2020;270:78-82. doi:10.3233/SHTI200126.

[10] ISO 13606 standard. https://www.iso.org/standard/67868.html. Accessed January 7, 2022.

[11] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform*. 2006;121:279-290.

[12] McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem*. 2003;49(4):624-633. doi:10.1373/49.4.624

[13] STOP-CORONAVIRUS. https://imas12.es/blog/stop-coronavirus-nuevo-proyecto-clinico-llevado-a-cabo-en-el-instituto-i12-para-ofrecer-respuestas-integrales-a-la-covid-19/. Accessed January 7, 2022.

[14] Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. doi:10.1016/j.jbi.2019.103208R.

[15] ISARIC-WHO CRF for COVID-19. https://isaric.org/research/covid-19-clinical-research-resources/covid-19-crf/. Accessed January 7, 2022.

[16] ISARIC Clinical Characterisation Group. The value of open-source clinical science in pandemic response: lessons from ISARIC [published correction appears in Lancet Infect Dis. 2021 Dec;21(12):e363]. *Lancet Infect Dis*. 2021;21(12):1623-1624. doi:10.1016/S1473-3099(21)00565-X.

[17] Foundation for Statistical Computing. https://www.r-project.org/about.html. Accessed January 7, 2022.

[18] Maldonado JA, Marcos M, Fernández-Breis JT, et al. CLIN-IK-LINKS: A platform for the design and execution of clinical data transformation and reasoning workflows. *Comput Methods Programs Biomed*. 2020;197:105616. doi:10.1016/j.cmpb.2020.105616.

[19] Ong TC, Kahn MG, Kwan BM, et al. Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. *BMC Med Inform Decis Mak*. 2017;17(1):134. Published 2017 Sep 13. doi:10.1186/s12911-017-0532-3.