Interoperable and interpretable Machine Learning based Data quality evaluation in obstetrics real-world data

João Coutinho-Almeida[1,2][0000−0003−0882−6547], Carlos Saez[], Ricardo João Cruz-Correia[1,2][0000−0002−3764−5158], and Pedro Pereira Rodrigues,[1,2][0000−0001−7867−6682]

[1] CINTESIS - Centre for Health Technologies and Services Research, University of Porto, Portugal

[2] MEDCIDS – Faculty of Medicine of University of Porto, Portugal

Abstract Keywords: Data Quality · Machine-learning · Bayesian Networks · Real-world data · FHIR

1 Introduction

With the wide spreading of healthcare information systems across all contexts of healthcare practice, the production of health-related data has followed this incremental behaviour. The potential for using this data to create new clinical knowledge and push medicine further is tempting [1]. However, to correctly use the data stored in Electronic health records (EHRs), the quality of the data must be robust enough to sustain the clinical decisions made based on this data. The issue is that data quality is not a straight line and is very context aware. The threshold and dimensions required to classify the quality of the data depend on the purpose that we intend to use that very same data [2]. These uses can be very distinct and have different impacts as well. For one, we can use data to support day-to-day decisions regarding individual patients' care [3]. These decisions can include ones based on recorded information to understand a patient's history, clinical decision support systems based on this data, or even using the data to help support a more macro, public health-oriented decision. Another area is using information for management purposes. The data can be used by management bodies and regulatory authorities to extract metrics regarding the quality of care or reimbursement purposes. Thirdly, data can be used for research purposes, namely observational studies and, more recently, to support clinical trials through real-world evidence analysis [4,3,5]. So, all the EHR data-based decisions can only be as good as the data supporting them. Several studies have already warned about the lack of data quality in EHRs and how this can be a significant hurdle to an accurate representation of the population and potentially lead to erroneous healthcare decisions [6,7,8,9,10,11].

There are several steps in the data lifecycle that can be prone to error, from data generation, where the data is registered by healthcare professionals, passing by data processing, whether inside healthcare institutions or by software engineers aiming to reuse data, to data interpretation and reuse, where investigators try to interpret the meaning of registered data [5]. So, with all of the data's possible uses added to the several steps that can introduce errors throughout the data lifecycle, data quality frameworks and sequential implementations can have very distinct approaches and methodologies to assess data quality. Data quality tools for checking data being registered live to support day-to-day decisions will be significantly different from one whose only purpose is to provide quality checks for research purposes. So, methodologies to tackle these issues are necessary for guaranteeing the quality of healthcare practice and the knowledge derived from EHR data. Consequently, in this paper, we propose:

- Create a tool for identifying data quality issues in obstetrics EHRs;
- Enlighten on the issues that can appear with a full deployment of such a tool;
- Suggestion of a creation of a single score for data quality for comparison of high-quality and low-quality records in a database.
- Assess how such a tool can work in early-stage real-world scenarios and how to work with obstetricians to improve data quality.
- Identify data quality issues on obstetrics data

2 Background and Related Work

There is already a significant number of papers trying to define data quality assessment frameworks for EHR data, all of them plausible and recommendable, already described in other papers [12]. The literature has over 20 different methods, descriptions, and summaries of different frameworks over the years. Some may be highlighted from the review from Weiskopf et. al, [13], where five data quality concepts were identified over 230 papers: Completeness, Correctness, Concordance, Plausibility, and Currency.

Then Khan et al. tried to harmonize data quality assessment frameworks, which simplified all previous concepts into three main categories: Conformance, Completeness, and Plausibility, and two assessment contexts: Verification and Validation [14]. Then a review of Bian et al. [12] expanded on the previous ones, categorizing data quality into 14 dimensions and mapping them to the previous most known definitions. These were: currency, correctness, plausibility, completeness, concordance, comparability, conformance, flexibility, relevance, usability, security, information loss, consistency, and interpretability. Finally, the work of Saez et. al., defined a unified set of DQ dimensions: completeness, consistency, duplicity, correctness, timeliness, spatial stability, contextualization, predictive value, and reliability [15]. Despite all of these comprehensive works, there is still no consensus regarding which one is best or which has taken the lead in usage. Moreover, looking at all of the descriptions related in the literature, a significant portion of concepts are overlapping, and sometimes hard to conceptualize such dimensions in practice.

As for implementations, there are already some available, such as the work from [16] where a tool created by primary care in the Flanders was built to assess completeness and percentage of values within the normal range. The work from Liaw et al. [17] already reviewed some data quality assessment tools, like tools from OHDSI [18] or TAQIH [19]. Additionally, we found some others with similar purposes and characteristics like the work presented data dataquieR [20], an R language-based package that can assess several data quality dimensions in observational health research data. Also, the work from Razzaghi et al. developed a methodology for assessing data quality in clinical data [21], taking into account the semantics of data and their meanings within their context. Furthermore, the work from Rajan et al. [22] presented a tool that can assess data quality and characterize health data repositories. Parallel to this, Kaspner et al. created a tool called DQAStats that enables the profiling and quality assessment of the MIRACUM database, being possible to integrate into other databases as well [23].

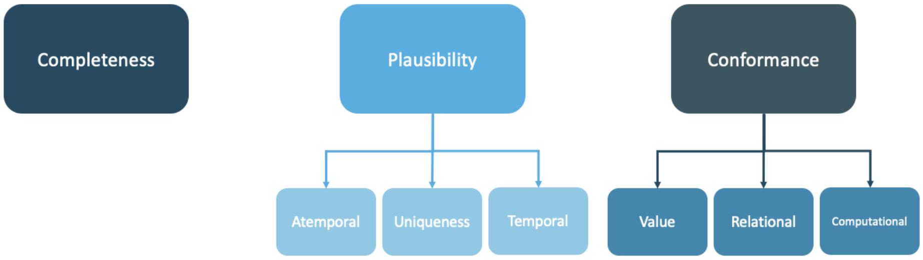
Regarding data quality assessment as a whole, the works of [24], focused on outlier detection in large-scale data repositories. The works of [25] focused on the exploration and identification of data-set shifts, contributing to the broad examination and repurposing of large, longitudinal data sets. Finally, the work of [26] focused on the manipulation of EHR data, including data quality assessment, data cleaning, and data extraction. However, these tools are not meant to be used at the production level, assessing data as it is being registered or outputs reports for human consumption and not a quantitive metric for metric comparison. Furthermore, none of the non-agnostic tools were designed for obstetric EHR data nor had interoperability based on standards in mind. Finally, we have not seen, until the moment of this paper, any implementation that used machine learning to evaluate the correctness of the value.

3 Materials

The data was gathered from 9 different Portuguese hospitals regarding obstetric information: data from the mother, several data points about the fetus and delivery mode. The data is from 2019 to 2020. The software for collecting data was the same in every institution, and the columns were the same, even though the version of each software differed across hospitals. Across the different hospitals, data rows ranged from 2364 to 18177. The sum of all rows rendered 73351 rows. The data dictionary is in appendix A.

For this purpose, we took the Khan harmonized framework since we understood it as simpler to communicate we feel that the three main categories are indeed non-reducible, which makes sense from an organizational standpoint. Furthermore, the work done by Khan et al. with mapping to already existing frameworks could help compare this work with others that felt the need to use other frameworks. With this in mind, we will use three main categories, Completeness, Conformance and Plausibility. Completeness relates to missing data. Conformance relates to the compliance of the data representation, like formatting, computational conformance and other data standards implemented. Plausibility relates to how believable the values are.

Figure 1. Dimensions of data quality



4 Methods

We wrote all of the code in Python 3.10.6 with the usage of the scikit-learn library for preprocessing, and evaluation [27]. For plausibility, a Bayesian network was used. We used this model due to the possiblity of using a single model for classifying the plausibility of all columns and due to its interpretable nature. The networks was created with the pgmpy package [28]. For creating the network, all null representations were standardized. Data was prepossessed with the removal of features with high missing rates ($> 80\%$ overall). All missing value representations were standardized. The imputation process was performed with the median for continuous and a new category (NULLIMP) for categorical variables. Then, the continuous variables were discretized into three bins defined by quantile. The evaluation was done with cross-validation with 10 splits and two repetitions for each column as the target.

As for Z-Scores, they were defined for all continuous variables based on the interquartile range. Then, rows were also assessed with distance analysis, with Local Outlier Factor and Elliptic Envelope from scikit-learn and the outlier-tree algorithm. We also added a rule engine, using the *great_expectations* package. Rules were defined by the team, focusing on impossible numbers present in age, weight, or relationship between variables. As for missing information was created with all the data, creating the scoring based on the inverse of the missing percentage. Missing detection was based on primary key variables. For completeness, we used the inverse of the percentage of nulls in the training set. The API for serving the prediction models was developed with FastAPI. So, the methods applied in terms of the DQA framework shown in figure 1 are described in the table 1.

The method of scoring was to obtain a single value that could grasp the quality of the row or patient. To assess the tool's usefulness, we will implement

Table 1. Implemented Methods

Category	Subcategory	Method
Completeness	N/A	Score by the inverse percentage of missing in the train data
Plausibility	Atemporal Plausibility	Bayesian model prediction based on the other values of row
Plausibility	Atemporal Plausibility	Z-score for column value based on IQR train data
Plausibility	Atemporal Plausibility	Elliptic Envelope
Plausibility	Atemporal Plausibility	Local Outlier Factor
Conformance	Value Conformance	Manual Rule engine
Plausibility	Atemporal Plausibility	Manual Rule engine
Plausibility	Atemporal Plausibility	outlier-tree

it in a production environment and collect metrics regarding the data being produced. Then we intended to present some results to selected obstetrics clinicians for them to assess how likely the information is to be suitable for usage. We will also compare the results with the ones from the model to make sanity checks regarding the model's performance and adequacy. The metrics of agreements will be the Cohen kappa and a ranking metric - Normalized Discounted Cumulative Gain (NDCG) [29] which is the sum of the true scores ranked in the order induced by the predicted scores, after applying a logarithmic discount. Then divide by the best possible score to obtain a score between 0 and 1.

5 Results

A Bayesian network with structure and parameters learned from the training dataset reached an average of Area Under the Receiver Operating Characteristic Curve of 0.857. The results are in the table 2.

The network is as represented in figure 2.

As for the rules created, they were conformance based, like the format of dates, and conformance to the value set (i.e. Robson group, bishop scores, or delivery types). We also added plausibility rules, like expected values for BMI, weight and gestational age. We also added plausibility for the relationship between columns, namely weight across different weeks of gestation. We have also added a relationship of greatness between ecography weights more than 5 weeks apart.

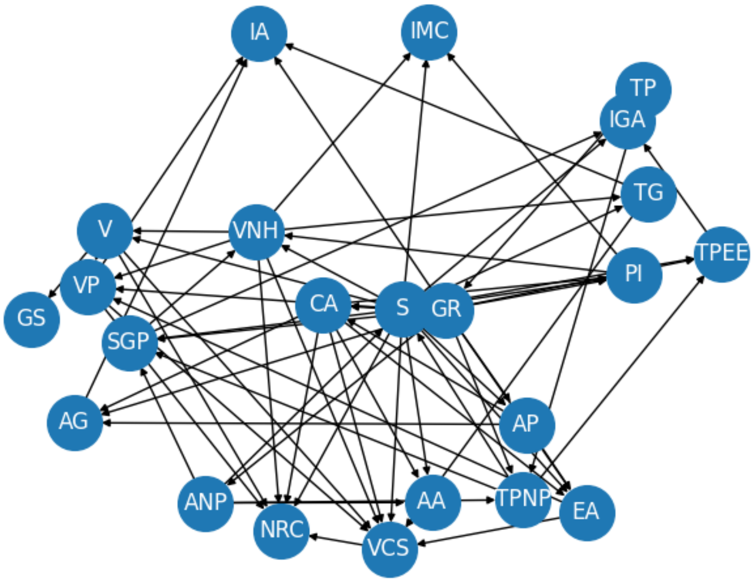
5.1 Deployment & Validation

The purpose of this model is to be served as an API for usage within a healthcare institution and act as a supplementary decision support tool for obstetrics

Table 2. Validation Results: Column acronym with AUROC along with 95% CI

AP	0.944	[0.943, 0.945]	VNH	0.894	[0.893, 0.895]
AG	0.797	[0.778, 0.816]	TPEE	0.816	[0.815, 0.816]
EA	0.969	[0.968, 0.969]	AA	0.751	[0.743, 0.758]
CA	0.958	[0.958, 0.958]	GR	0.931	[0.93, 0.932]
IA	0.638	[0.637, 0.638]	V	0.983	[0.982, 0.983]
PI	0.881	[0.88, 0.881]	TP	0.866	[0.865, 0.868]
IMC	0.881	[0.881, 0.882]	VCS	0.79	[0.789, 0.791]
NRC	0.75	[0.75, 0.75]	ANP	0.942	[0.938, 0.946]
IGA	0.968	[0.968, 0.969]	GS	0.514	[0.507, 0.52]
SGP	0.974	[0.974, 0.974]	S	0.896	[0.896, 0.897]
VA	0.974	[0.974, 0.974]	VP	0.771	[0.77, 0.772]
TG	0.728	[0.726, 0.73]	TPNP	0.952	[0.951, 0.952]
Average 0.857 [0.846, 0.868]					

Figure 2. Network learned



teams. Although a concrete, vendor-specific information model and health information system were initially used, our goal is to develop a more universal clinical decision support system. This system should be usable across all systems involved in birth and obstetrics departments. Therefore, we constructed it using the Health Level 7 (HL7) Fast Healthcare Interoperable Resources (FHIR) R5 version standard. This approach simplifies the process of API interaction. Rather than utilizing a proprietary model for the data, we based our decision on the use of FHIR resources: Bundle and Observation. These resources handle the request and response through a customized operation named "$quality_check". Our intention is to publish the profiles of these objects to streamline API access via standardized mechanisms and data models. The current version of the profiles can be accessed at this URL: https://joofio.github.io/obs-cdss-fhir/.

For validation, we deployed the tool in docker format in a hospital to gather new data. We gathered 3231 new cases and returned a score for quality as exemplified in figure 3. Being that the score is from 0 to 1, the average score was 0.23 and IQR was 0.03. We also used the clinician from one of the hospitals that we get data from and asked this clinician to assess 10 records in terms of quality. We gathered the 10 records at random and asked the clinician to assess them in terms of quality. Our purpose was then to compare the rankings of each evaluator; the model and the clinician, in order to assess how similar they were as can be seen in figure 4.

Figure 3. Model score for newly seen data

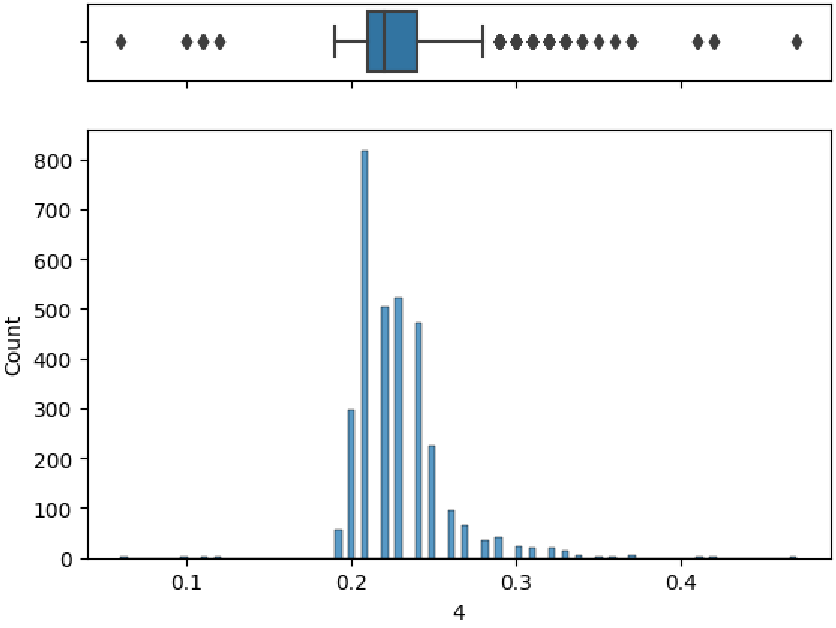
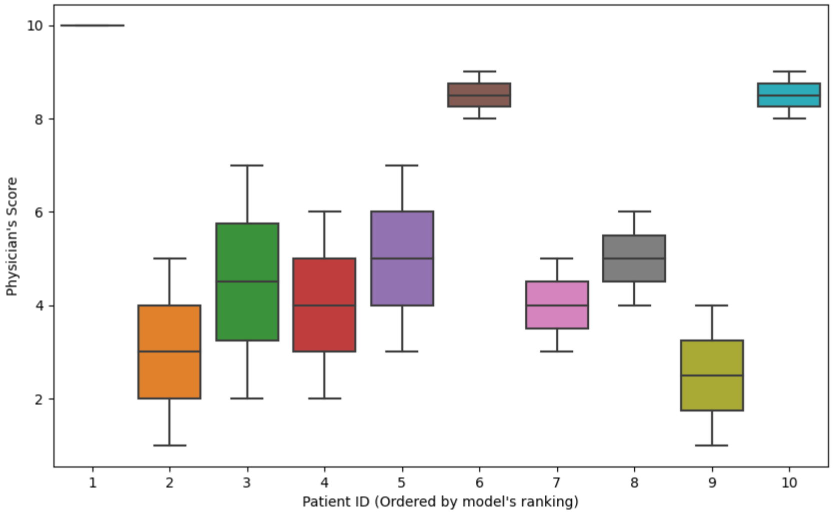


Figure 4. Comparasion of clinical assessment of records with the model



6 Discussion

The first thing to address is that data quality is still an elusive concept since it has a contextual dimension and the quality of the record depends on the usage of the information. For example, data aimed at primary usage and day-to-day healthcare decisions about a patient will have different requirements regarding the importance of some variable or completeness of information very different from data needed to create summary statistics for key performance indicators extraction. Moreover, the data is still very vendor-specific. Even though we used a interoperability standards, the semantic layer, more connected with terminology is still lacking. This is a issue to be addressed in order to improve the interoperability of the standard. Moreover, we do not know how the training done with this data is generalizable to other vendors. One opportunity arises of mapping all of this data to a widely used terminology like snomed or loinc.

7 Conclusion

This work is still an early draft of a production-ready tool. However, we feel the work done is already a valuable insight into how to use data quality frameworks and several statistical tools in order to assess ehr data quality. This is a fundamental process not only to guarantee the quality of data for primary usage on a day-to-day but also for securing quality for secondary analysis and usage. We believe the fact that we created an interoperable tool that was trained on real obstetrics data from 9 different hospitals and has the ability to provide a single score for a clinical record can help institutions, academics, and ehr vendors implement data quality assessment tools in their own systems and institutions.

For the next steps, we would like to further evaluate the score and its relationship with clinical usefulness. This would also include a further assessment of a threshold for the score for defining a record that would require human attention.

References

1. F. Martin-Sanchez and K. Verspoor. Big data in medicine is driving big changes. *Yearbook of Medical Informatics*, 9:14–20, August 2014.
2. Muhammad F. Walji. Electronic Health Records and Data Quality. *Journal of Dental Education*, 83(3):263–264, March 2019.
3. Robert A. Verheij, Vasa Curcin, Brendan C. Delaney, and Mark M. McGilchrist. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *Journal of Medical Internet Research*, 20(5):e9134, May 2018.
4. Kristin M. Corey, Joshua Helmkamp, Morgan Simons, Lesley Curtis, Keith Marsolo, Suresh Balu, Michael Gao, Marshall Nichols, Joshua Watson, Leila Mureebe, Allan D. Kirk, and Mark Sendak. Assessing Quality of Surgical Real-World Data from an Automated Electronic Health Record Pipeline. *Journal of the American College of Surgeons*, 230(3):295–305.e12, March 2020.
5. Chunhua Weng. Clinical data quality: A data life cycle perspective. *Biostatistics & Epidemiology*, 4(1):6–14, January 2020.
6. Andrew P. Reimer, Alex Milinovich, and Elizabeth A. Madigan. Data quality assessment framework to assess electronic medical record data for use in research. *International Journal of Medical Informatics*, 90:40–47, June 2016.
7. Erik Joukes, Nicolette F. de Keizer, Martine C. de Bruijne, Ameen Abu-Hanna, and Ronald Cornet. Impact of Electronic versus Paper-Based Recording before EHR Implementation on Health Care Professionals' Perceptions of EHR Use, Data Quality, and Data Reuse. *Applied Clinical Informatics*, 10(2):199–209, March 2019.
8. Vojtech Huser, Frank J. DeFalco, Martijn Schuemie, Patrick B. Ryan, Ning Shang, Mark Velez, Rae Woong Park, Richard D. Boyce, Jon Duke, Ritu Khare, Levon Utidjian, and Charles Bailey. Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets. *EGEMS (Washington, DC)*, 4(1):1239, 2016.
9. Yili Zhang and Güneş Koru. Understanding and detecting defects in healthcare administration data: Toward higher data quality to better support healthcare operations and decisions. *Journal of the American Medical Informatics Association: JAMIA*, 27(3):386–395, March 2020.
10. Oren Kramer, Adir Even, Idit Matot, Yohai Steinberg, and Yuval Bitan. The impact of data quality defects on clinical decision-making in the intensive care unit. *Computer Methods and Programs in Biomedicine*, 209:106359, September 2021.
11. Mark J. Giganti, Bryan E. Shepherd, Yanink Caro-Vega, Paula M. Luz, Peter F. Rebeiro, Marcelle Maia, Gaetane Julmiste, Claudia Cortes, Catherine C. McGowan, and Stephany N. Duda. The impact of data quality and source data verification on epidemiologic inference: A practical application using HIV observational data. *BMC public health*, 19(1):1748, December 2019.

12. Jiang Bian, Tianchen Lyu, Alexander Loiacono, Tonatiuh Mendoza Viramontes, Gloria Lipori, Yi Guo, Yonghui Wu, Mattia Prosperi, Thomas J. George, Christopher A. Harle, Elizabeth A. Shenkman, and William Hogan. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *Journal of the American Medical Informatics Association: JAMIA*, 27(12):1999–2010, December 2020.
13. Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, January 2013.
14. Michael G. Kahn, Tiffany J. Callahan, Juliana Barnard, Alan E. Bauck, Jeff Brown, Bruce N. Davidson, Hossein Estiri, Carsten Goerg, Erin Holve, Steven G. Johnson, Siaw-Teng Liaw, Marianne Hamilton-Lopez, Daniella Meeker, Toan C. Ong, Patrick Ryan, Ning Shang, Nicole G. Weiskopf, Chunhua Weng, Meredith N. Zozus, and Lisa Schilling. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs*, 4(1):1244, September 2016.
15. Carlos Sáez, Juan Martínez-Miranda, Montserrat Robles, and Juan Miguel García-Gómez. Organizing Data Quality Assessment of Shifting Biomedical Data. 2012.
16. Hang T. T. Phan, Florina Borca, David Cable, James Batchelor, Justin H. Davies, and Sarah Ennis. Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: Protocol and application to a large patient cohort. *Scientific Reports*, 10(1):10164, June 2020.
17. Siaw-Teng Liaw, Jason Guan Nan Guo, Sameera Ansari, Jitendra Jonnagaddala, Myron Anthony Godinho, Alder Jose Borelli, Simon de Lusignan, Daniel Capurro, Harshana Liyanage, Navreet Bhattal, Vicki Bennett, Jaclyn Chan, and Michael G. Kahn. Quality assessment of real-world data repositories across the data life cycle: A literature review. *Journal of the American Medical Informatics Association: JAMIA*, 28(7):1591–1599, July 2021.
18. George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R. Rijnbeek, Johan van der Lei, Nicole Pratt, G. Niklas Norén, Yu-Chuan Li, Paul E. Stang, David Madigan, and Patrick B. Ryan. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in Health Technology and Informatics*, 216:574–578, 2015.
19. Roberto Álvarez Sánchez, Andoni Beristain Iraola, Gorka Epelde Unanue, and Paul Carlin. TAQIH, a tool for tabular data quality assessment and improvement in the context of health data. *Computer Methods and Programs in Biomedicine*, 181:104824, November 2019.
20. Carsten Oliver Schmidt, Stephan Struckmann, Cornelia Enzenbach, Achim Reineke, Jürgen Stausberg, Stefan Damerow, Marianne Huebner, Börge Schmidt, Willi Sauerbrei, and Adrian Richter. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC medical research methodology*, 21(1):63, April 2021.
21. Hanieh Razzaghi, Jane Greenberg, and L. Charles Bailey. Developing a systematic approach to assessing data quality in secondary use of clinical data based on intended use. *Learning Health Systems*, 6(1):e10264, 2022.
22. Naresh Sundar Rajan, Ramkiran Gouripeddi, Peter Mo, Randy K. Madsen, and Julio C. Facelli. Towards a content agnostic computable knowledge repository

for data quality assessment. *Computer Methods and Programs in Biomedicine*, 177:193–201, August 2019.

23. Lorenz A. Kapsner, Jonathan M. Mang, Sebastian Mate, Susanne A. Seuchter, Abishaa Vengadeswaran, Franziska Bathelt, Noemi Deppenwiese, Dennis Kadioglu, Detlef Kraska, and Hans-Ulrich Prokosch. Linking a Consortium-Wide Data Quality Assessment Tool with the MIRACUM Metadata Repository. *Applied Clinical Informatics*, 12(4):826–835, August 2021.

24. Hossein Estiri and Shawn N Murphy. Semi-supervised Encoding for Outlier Detection in Clinical Observation Data. *Computer methods and programs in biomedicine*, 181:104830, 2019.

25. Carlos Sáez, Alba Gutiérrez-Sacristán, Isaac Kohane, Juan M García-Gómez, and Paul Avillach. EHRtemporalVariability: Delineating temporal data-set shifts in electronic health records. *GigaScience*, 9(8):giaa079, 2020.

26. David A. Springate, Rosa Parisi, Ivan Olier, David Reeves, and Evangelos Kontopantelis. rEHR: An R package for manipulating and analysing Electronic Health Record data. *PloS One*, 12(2):e0171784, 2017.

27. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

28. Ankur Ankan and Abinash Panda. pgmpy: Probabilistic graphical models using python. In *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*. Citeseer, 2015.

29. Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, and Wei Chen. A theoretical analysis of ndcg type ranking measures, 2013.

A Data Dictionary

Initial	Description
IA	Mother Age
GS	Blood Group
PI	Weight at the beginning of pregnancy
PAI	Weight on Admission
IMC	BMI
CIG	If Smoker During Pregnancy
APARA	Number of previously born babies
AGESTA	Number of Pregnancies
EA	Number of Previous Eutocic Deliveries with no assistance
VA	Number of Previous Eutocic Deliveries with help of vacuum extraction
FA	Number of Previous Eutocic Deliveries with help of forceps
CA	Number of Previous C-sections
TG	Pregnancy Type (spontaneous, In vitro fertilisation...)
V	If the pregnancy was accompanied by physician
NRCPN	Number of prenatal consultations
VH	If the pregnancy was accompanied by a physician in a hospital
VP	If the pregnancy was accompanied by a physician in a private clinic
VCS	If the pregnancy was accompanied by a physician in a primary care facility
VNH	If the pregnancy was accompanied by a physician in the hospital the delivery was made
B	Pelvis Adequacy
AA	Baby's Position on Admission
BS	Bishop Score
BC	Bishop Score Cervical Consistency
BDE	Bishop Score Fetal Station
BDI	Bishop Score Dilatation
BE	Bishop Score Effacement
BP	Bishop Score Cervical Position
IGA	Number of Weeks on Admission
TPEE	If the delivery was spontaneous
TPEI	If the delivery was induced
RPM	If there was a rupture of the amniotic pocket before delivery began
DG	Gestational Diabetes
TP	Delivery Type
ANP	Baby's Position on Delivery
TPNP	Actual Type of Delivery
SGP	Pregnancy Weeks on Delivery
GR	Robson Group