Characterizing Publicly Available Tabular Health Data Sets for Responsible Machine Learning

Mariana Oliveira[1,2] (✉)[0000−0001−9708−1123] and Carlos Soares[1,2,3]

[1] Faculty of Engineering - University of Porto, Porto, Portugal
mariana.oliveira@fe.up.pt
[2] Artificial Intelligence and Computer Science Laboratory (LIACC)
[3] Fraunhofer Portugal AICOS

Abstract. Publicly available data sets play a crucial role in reliable, reproducible and responsible Machine Learning (ML) research and application. However, common benchmark data sets often suffer from data quality and other issues that affect model behavior. Auditing public datasets is, thus, a crucial step in understanding data, allowing better models to be developed and improving understanding of their results. In this paper, we characterize popular healthcare data sets to illustrate how this can help practitioners select benchmarking data that align with their needs for responsible development of ML models. We conduct our study using 15 popular healthcare benchmarking data sets available on the UCI Machine Learning Repository. To assess their quality, we consider measures of accuracy, completeness and redundancy. To assess their complexity, we consider feature overlap, linearity, neighborhood, network, dimensionality, and class balance measures. We also inspect the data sets for imbalanced representation of different demographic groups. We found several data quality and demographic representation imbalance issues that indicate that careful analysis of public data sets remains needed.

Keywords: Responsible ML; Data auditing; Public healthcare data

1 Introduction

Publicly available data sets play a critical role in the development and dissemination of ML research and application. The lack of access barriers makes these data sets appealing for preliminary testing, but also for transparent sharing of results. This can ensure not only scientific reproducibility, but also accountability and auditability – a key dimension of responsible and trustworthy AI [23].

Responsible ML development also requires robustness to data quality issues, which should be evaluated quantitatively. Auditing data quality is, thus, a crucial step in ML pipelines, especially in applications like healthcare where the potential human impact is high. Several methodologies for data quality assessment have been proposed in the literature [6]. Different methodologies often propose different *data quality dimensions* that should be considered, as well as measures to assess them. Recent proposals extend data quality dimensions to consider bias and fairness [31], since this is another key dimension of responsible ML [23].

Data complexity can also increase the difficulty of developing responsible ML solutions. More complex data may require more complex models, which can reduce explainability. Initially proposed for classification problems [19], data complexity measures, such as feature overlap and separability of classes, have been extended to address other tasks and challenges (e.g., [26,5]).

Investigating data quality and complexity using such measures can help to recognize the challenges posed by certain data sets. This can improve understanding of the strengths and limitations of a certain ML approach, leading to more responsible use of ML. Many papers assess data set quality and other data properties, either as part of data quality assessment and exploration steps in a standard ML pipeline, or to use them as meta-features for meta-learning tasks [33,2]. However, we are unaware of a recent study focusing on the data quality and complexity of publicly available data sets in the healthcare domain.

The UCI Machine Learning repository [15] is one of the most famous in the ML community, hosting over 600 easily downloadable public data sets, 109 of which related to Health and Medicine. Other repositories, such as PhysioNet [17], focus specifically on health data and sometimes require training or authorization before having access to certain data sets.

In this paper, we aim to inspect the quality and characterize some of the most popular "Health and Medicine" data sets currently available in the UCI Machine Learning repository. In Section 2, we discuss related work. In Section 3, we present the selected data sets and data quality and complexity measures computed. In Section 4, we present and discuss our results. We conclude with some insights about responsible application of ML in healthcare in Section 5.

2 Related Work

Recently, Longjohn et al. [24] highlighted how data quality issues affect common benchmark datasets, including illustrative examples from the UCI ML repository. Kohli et al. [21] found that most tabular datasets used to benchmark ML approaches are also quite old, potentially affecting results. However, these studies do not provide detailed measures of data set quality and complexity.

In 2014, Macià & Bernadó-Mansilla [27] evaluated the quality and complexity of data sets available on the UCI ML repository at the time. However, the scope of their work was more general and only 2 of the 15 datasets in this paper were included. We also compute some additional data quality and complexity measures, and investigate the distribution of demographic variables in each dataset.

3 Data and Methods

3.1 Data Quality Measures

Batini et al. [6] suggest considering accuracy, completeness, redundancy, readability, accessibility, consistency, usefulness, and trust as data quality dimensions. Each dimension may be quantitatively assessed using one or more measures. When working with public data sets, some dimensions can be easily and

Table 1: Data quality measures computed. These are based on the tabular input data quality measures available in Python library *pymdma* [12,13].

accuracy [4]	*rare*: % of rare ($freq < 5\%$) nominal values in cat. variables
	$outIQR$: % of outliers in num. variables according to the boxplot rule
completeness	$incR/incC$: % of incomplete rows/columns
	$missR/missC$: average % of missing values per row/column
redundancy	*dupl*: % of duplicated rows
	$strCorr$: % of num. variable pairs with absolute correlation above 0.5
	corr: average and standard deviation of correlation values

accurately assessed (e.g., calculating the percentage of null values as a measure of *completeness*); others, such as timeliness may not apply.

Many software tools for data quality assessment exist [11], with some focusing specifically on the healthcare domain [34]. We based our work on the Python library ***pymdma**: multimodal data measures for auditing real and synthetic datasets* [12,13]. Using the *R* language, we re-implemented some of the tabular input data quality measures described in [4], with some modifications. A list of the measures we computed can be found in Table 1.

Data may also be affected by different types of bias [31], which can be viewed as affecting data quality. In this study, we investigate the imbalance in representation of different demographic groups in the data sets. Being aware of imbalanced representation in data is a step towards being able to properly assess and mitigate bias that may contribute to a lack of fairness in an ML solution.

3.2 Data Complexity Measures

Lorena et al. [25] describe in detail the classification complexity measures in Table 2, which were defined so that higher values indicates higher complexity.[5] We computed these measures using the implementation available in *R* package ***ECoL*** [16,26,25]. The measures can only be computed for complete data sets, and categorical variables are internally binarized. Before calculating them, we pre-processed the data sets by excluding any feature with more than 20% missing values, and any categorical feature with more than 50 unique nominal values.

All code needed to reproduce this study is available on Github.[6]

3.3 Data Sets

Fifteen data sets were selected from the UCI Machine Learning Repository [15]. Filtering by subject area "Health and Medicine" and data type "Multivariate" or "Tabular", we considered the top 15 data sets in terms of number of views [7] that also met the conditions detailed next.

[4] While rare values may be accurate, they should be checked for potential errors.

[5] The exception was the entropy of class proportions (*B1*) where a higher value corresponded to a more balanced (assumed to be simpler) problem. Here, we will take the complement of that measure so that a high value indicates high complexity.

[6] https://github.com/mrfoliveira/Characterizing-public-healthcare-data-RHCML25

[7] The number of views of the selected data sets ranged between 27.24k and 790.87k.

Table 2: Complexity measures computed. These measures are described in [25] and available in *R* package *ECoL* [16,26,25].

feature-based	$F1/F1v$ Fisher's discriminant ratio, and its directional-vector
	$F2$: overlapping of the per-class bounding boxes
	$F3/F4$: maximum individual/collective, feature efficiency
linearity	$L1$: distance of erroneous instances to a linear classifier
	$L2$: training error of a linear classifier
	$L3$: nonlinearity of a linear classifier
neighborhood	$N1$: fraction of points lying on the class boundary
	$N2$: average intra/inter class nearest neighbor distances
	$N3$: leave-one-out error rate of the 1-NN algorithm
	$N4$: nonlinearity of the 1-NN classifier
	$T1$: fraction of maximum covering spheres on data
	LSC: local-Set cardinality
network	$Density$: density of network
	$ClsCoef$: clustering Coefficient
	$Hubs$: hub score
dimensionality	$T2$: number of samples per dimension
	$T3$: intrinsic dimensionality per number of examples
	$T4$: intrinsic dimensionality proportion
class label balance	$B1$: entropy of class proportions
	$B2$: multi-class imbalance ratio

Data sets were only considered if they: were available for download through the *R* package ***ucimlrepo*** [3]; were not derived from others already included; had 100 or more instances, a single variable tagged as target, and at least one demographic variable identified in the metadata. Variables tagged with the role "ID" or "Other" in the metadata and those with a single unique value were removed from the data sets and excluded from this analysis. Demographic variables in the data represent potentially sensitive information regarding age (A), gender (G), sex (S), race (R), sexual orientation (SO), education level (EL) and income (I). The data sets were all mainly geared towards classification, though some include variables that could potentially be used as targets for a regression task. The basic data characteristics after these exclusions can be found in Table 3.

We found some issues with the metadata of some of the selected data sets, related to the role and type of variables included, and missing values encoding (details available in supplementary material on Github).

4 Results and Discussion

Results are shown in Figures 1 and 2. In both figures, data sets with similar scores appear together after hierarchical clustering using Euclidian distance.

In Figure 1, we can see that 8 of the data sets had no missing values. Only 6 of the data sets show a meaningful percentage of numeric outliers and/or rare nominal values. Data set $D9$ had the highest redundancy, with a rather high percentage of duplicate values; it is joined by data set $D3$ as the only two data sets with at least 10% of their feature pairs showing a strong correlation.

In Figure 2, we can see the target variable is quite balanced in most data sets; with data sets $D4$, $D1$, and $D12$ standing out as the most imbalanced. On average, measures related to class balance and feature overlap (except *F1.mean*)

Table 3: Basic characteristics of the data sets, including the number of columns excluded for (a) having role "Other" or "ID", or (b) having a single unique value. Note that the target also did not count towards the categorical features total.

ID	Name	#Class	#Inst.	#Cols.	Feats.	#Num.	#Cat.	#Dem.	Dem.	Excl.(a)	(b)
D1	National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset [29]	2	2.28k	8	Mixed	6	1	2	A, G	2	-
D2	AIDS Clinical Trials Group Study 175 [18]	2	2.14k	23	Mixed	12	10	4	A, G, R, SO	1	1
D3	ILPD (Indian Liver Patient Dataset) [32]	2	583	11	Mixed	9	1	2	A, G	-	-
D4	CDC Diabetes Health Indicators [8]	2	254k	22	Mixed	7	14	4	A, EL, I, S	1	-
D5	Heart Failure Clinical Records [9]	2	299	13	Mixed	7	5	2	A, S	-	-
D6	Breast Cancer [37]	2	286	10	Mixed	1	8	1	A	-	-
D7	Differentiated Thyroid Cancer Recurrence [7]	2	383	17	Mixed	1	15	2	A, G	-	-
D8	Glioma Grading Clinical and Mutation Features [36]	2	839	24	Mixed	1	22	3	A, G, R	1	-
D9	Maternal Health Risk [1]	3	1k	7	Num.	6	0	1	A	-	-
D10	Cirrhosis Patient Survival Prediction [14]	3	418	18	Mixed	10	7	2	A, S	2	-
D11	Diabetes 130-US Hospitals for Years 1999-2008 [35]	3	102k	46	Mixed	8	37	3	A, G, R	2	2
D12	HCV data [22]	5	615	13	Mixed	11	1	2	A, S	1	-
D13	Heart Disease [20]	5	303	14	Mixed	6	7	2	A, S	-	-
D14	Estimation of Obesity Levels Based On Eating Habits and Physical Condition [30]	7	2.1k	17	Mixed	8	8	2	A, G	-	-
D15	National Poll on Healthy Aging (NPHA)[28]	3	714	14	Cat.	0	13	3	A, G, R	-	1

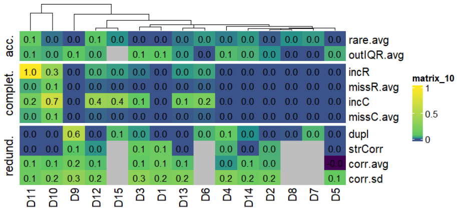

Fig. 1: Quality measures computed for the chosen data sets.

present the highest variance across data sets, followed by the variance of neighborhood measures [8]. For the latter two dimensions, it is possible to find data sets with very low (values close to 0) and very high complexity (values close to 1) according to different measures. The values for linearity show the least variance and relatively low values which would indicate less complexity.

Figure 3 shows the results of inspecting demographic variables in the data sets [9]. Age is the most frequently available variable. The more balanced variables tend to represent gender, while several variables representing race are the most imbalanced. Note that this analysis simply looks at the imbalance in the distribution of demographic variables in the data sets, and does not take into account the class labels within each demographic group.

Limitations Data quality issues and pre-processing decisions (e.g., about how to encode nominal variables with inherent order) may have somewhat distorted certain measures, especially those relating to complexity as both incomplete features and instances had to be discarded.

[8] Note that, due to computational resource limitations, it was not possible to compute some complexity measures for the two largest data sets, $D4$ and $D11$.

[9] Numeric variables were discretized to keep consistency with the measures calculated for categorical variables; we used 5 equal-width bins.

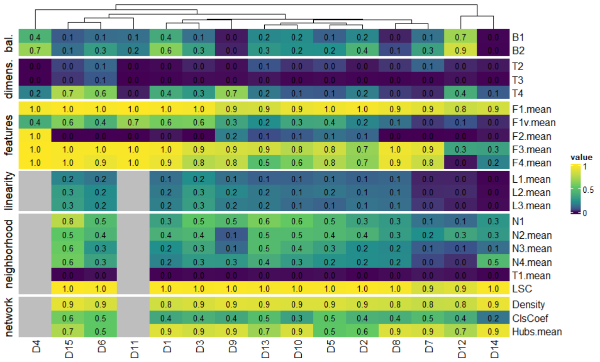

Fig. 2: Complexity of chosen data sets (higher values mean higher complexity).

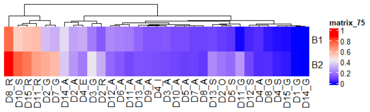

Fig. 3: Demographic imbalance measures (higher values mean greater imbalance).

5 Generalizable Insights about Responsible Application of Machine Learning in Healthcare

This work is a step towards better understanding of publicly available benchmarking data sets in the healthcare domain through available data quality and complexity assessment tools. We provide results that may help practitioners choose benchmark data sets that align with their responsible ML research goals.

Robustness and fairness are two key dimensions of responsible ML [23]; selecting appropriate data sets can help when checking that ML solutions follow these principles. A researcher might select a complex data set with data quality issues to test the robustness of an approach to issues that commonly affect real-world health data sets. Detecting and mitigating bias is especially important in the healthcare domain where it can lead to differing patient outcomes across different demographic groups [10]. Thus, a researcher might select one of the benchmark data sets that we found to have imbalanced demographic representation (e.g., the Glioma Grading data set) to test the fairness of an approach.

Using public data can support robust, fair and auditable ML development. However, we also found several metadata and other unreported issues in the data sets that indicate that careful analysis of public data sets remains needed.

Future work could extend this study to: include (a) a broader sample of healthcare data sets; (b) more measures of data quality since some dimensions, like trust, were not assessed; and (c) additional measures related to responsible and trustworthy AI (e.g., relating to data privacy); and explore how class labels and data quality relate to protected features, potentially impacting fairness.

Acknowledgments. This work was partially funded by project AISym4Med (101095387). AISym4Med is funded by the European Union under the Horizon Europe Framework Programme Grant Agreement Nº: 101095387. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Health and Digital Executive Agency can be held responsible for them.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmed, M.: Maternal Health Risk (2020). https://doi.org/10.24432/C5DP5D
2. Alcobaça, E., Siqueira, F., Rivolli, A., Garcia, L.P., Oliva, J.T., De Carvalho, A.C.: MFE: Towards reproducible meta-feature extraction. J. Mach. Learn. Res. **21**(111), 1–5 (2020)
3. Balamuta, J.J., Truong, P.: Ucimlrepo: Explore UCI ML Repository Datasets (2024)
4. Barandas, M.: Input validation. https://pymdma.readthedocs.io/en/v0.1.9/tabular/input_val/
5. Barella, V.H., Garcia, L.P.F., de Souto, M.C.P., Lorena, A.C., de Carvalho, A.C.P.L.F.: Assessing the data complexity of imbalanced datasets. Inf. Sci. **553**, 83–109 (Apr 2021)
6. Batini, C., Scannapieco, M.: Data and Information Quality: Dimensions, Principles and Techniques. Data-Centric Systems and Applications (2016)
7. Borzooei, S., Briganti, G., Golparian, M., Lechien, J.R., Tarokhian, A.: Machine learning for risk stratification of thyroid cancer patients. Eur. Arch. Oto-Rhino-Laryngol. **281**(4), 2095–2104 (Apr 2024)
8. CDC: Diabetes Health Indicators (2017). https://doi.org/10.24432/C53919
9. Chicco, D., Giuseppe Jurman: Heart Failure Clinical Records (2020). https://doi.org/10.24432/C5Z89R
10. Chinta, S.V., Wang, Z., Palikhe, A., Zhang, X., Kashif, A., Smith, M.A., Liu, J., Zhang, W.: AI-driven healthcare: A review on ensuring fairness and mitigating bias. PLOS Digital Health **4**(5), e0000864 (May 2025)
11. Ehrlinger, L., Wöß, W.: A Survey of Data Quality Measurement and Monitoring Tools. Front. Big Data **5**, 850611 (Mar 2022)
12. Façoco, I.S., Rebelo, J., Matias, P., Bento, N., Morgado, A.C., Sampaio, A., Rosado, L., Barandas, M.: pyMDMA: Multimodal data metrics for auditing real and synthetic datasets. SoftwareX **31**, 102256 (2025)
13. Façoco, I.S., Matias, P., Rebelo, J.: Fraunhoferportugal/pymdma: Patch version 0.1.9. Zenodo (Mar 2025). https://doi.org/10.5281/ZENODO.15064792
14. Fleming, T.R., Harrington, D.P.: Counting Processes and Survival Analysis. John Wiley & Sons (2013)
15. Frank, A.: UCI machine learning repository. http://archive. ics. uci. edu/ml (2010)
16. Garcia, L., Lorena, A.: ECoL: Complexity Measures for Supervised Problems (2019)
17. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)

18. Hammer, S.M., Katzenstein, D.A., Hughes, M.D., Gundacker, H., Schooley, R.T., Haubrich, R.H., Henry, W.K., Lederman, M.M., Phair, J.P., Niu, M.: A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with A critical comparative study of liver patients from USA and INDIA4 cell counts from 200 to 500 per cubic millimeter. N. Engl. J. Med. **335**(15), 1081–1090 (1996)
19. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3), 289–300 (2002)
20. Janosi, A., William Steinbrunn, Matthias Pfisterer, Robert Detrano: Heart Disease (1989). https://doi.org/10.24432/C52P4X
21. Kohli, R., Feurer, M., Eggensperger, K., Bischl, B., Hutter, F.: Towards quantifying the effect of datasets for benchmarking. In: ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR) (2024)
22. Lichtinghagen, R., Frank Klawonn, Georg Hoffmann: HCV data (2020). https://doi.org/10.24432/C5D612
23. Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., Tang, J.: Trustworthy AI. ACM Trans. Intell. Syst. Technol. **14**(1), 1–59 (Feb 2023)
24. Longjohn, R., Kelly, M., Singh, S., Smyth, P.: Benchmark data repositories for better benchmarking. Adv. Neural Inf. Process. Syst. **37**, 86435–86457 (2024)
25. Lorena, A.C., Garcia, L.P., Lehmann, J., Souto, M.C., Ho, T.K.: A survey on measuring classification complexity. ACM Comput. Surv. (CSUR) **52**(5), 1–34 (2019)
26. Lorena, A.C., Maciel, A.I., de Miranda, P.B., Costa, I.G., Prudêncio, R.B.: Data complexity meta-features for regression problems. Mach. Learn. **107**, 209–246 (2018)
27. Macià, N., Bernadó-Mansilla, E.: Towards UCI+. Inf. Sci. **261**, 237–262 (Mar 2014)
28. Malani, P.N., Kullgren, J., Solway, E.: National Poll on Healthy Aging (NPHA), [USA], April 2017 (2019). https://doi.org/10.3886/ICPSR37305.V1
29. National Center For Health Statistics (NCHS) At The Centers For Disease Control And Prevention (CDC): National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset (2019). https://doi.org/10.24432/C5BS66
30. Palechor, F.M., De la Hoz Manotas, A.: Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data Br. **25**, 104344 (2019)
31. Quaresmini, C., Giuseppe Primiero: Data Quality Dimensions for Fair AI. In: Proc. 2nd Workshop on Fairness and Bias in AI Co-Located with ECAI (2024)
32. Ramana, B.V., Babu, M.S.P., Venkateswarlu, N.B.: A critical comparative study of liver patients from USA and India: An exploratory analysis. IJCSI **9**(3), 506 (2012)
33. Rivolli, A., Garcia, L.P., Soares, C., Vanschoren, J., De Carvalho, A.C.: Meta-features for meta-learning. Knowl.-Based Syst. **240**, 108101 (Mar 2022)
34. Sánchez, R.Á., Iraola, A.B., Unanue, G.E., Carlin, P.: TAQIH, a tool for tabular data quality assessment and improvement in the context of health data. Comput. Methods Programs Biomed. **181**, 104824 (2019)
35. Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., Cios, K.J., Clore, J.N.: Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. BioMed Res. Int. **2014**(1), 781670 (2014)
36. Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K., Krauze, A.V.: Hierarchical voting-based feature selection and ensemble learning model scheme for glioma grading with clinical and molecular characteristics. Int. J. Mol. Sci. **23**(22), 14155 (2022)
37. Zwitter, M., Milan Soklic: Breast Cancer (1988). https://doi.org/10.24432/C51P4M