Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260527

Health Data Set Bias Examination in the European Health Data Space

Evangelia Anna MARKATOU[1,2] and Catherine CHRONAKI[1]
[1]HL7 Europe, Brussels, Belgium
[2]TU Delft, Delft, Netherlands
ORCiD: C. Chronaki: 0000-0001-6638-8448, E. A. Markatou: 0009-0005-3902-977X

Abstract. Diversity in health datasets is a necessary, quantifiable property of inclusivity across key domains (e.g., demographic, socioeconomic, health, and environmental) that directly shapes how well research generalizes and how fair its impacts are. In its absence, interventions risk encoding bias and exacerbating health disparities. In this work, we outline how dataset diversity can be measured and how these measures can be surfaced in the EHDS via metadata. We argue that simple annotations are insufficient: structured interaction with the data owner is required to assess the utility of health datasets for specific research purposes.

Keywords. European Health Data Space (EHDS), Diversity and Bias, Interoperability Standards, Data Quality, HealthDCAT-AP, HL7 FHIR

1. Introduction

Excellent research is diverse and inclusive. When studies reflect broad representation of the population, they reduce bias and yield results that are robust and generalizable. Inclusion is not only ethical and fair—it is socially responsible for science. It also aligns with regulatory expectations (e.g., European Health Data Space (EHDS) [1]) to document and examine bias in datasets. Practically, diversity boosts robustness, making models and conclusions more resilient. In short, *dataset diversity is a fundamental quality parameter,* an essential component of a dataset utility label.

The EHDS regulation fosters a health-specific ecosystem to advance research and innovation by creating trusted, cross-border access to electronic health data. Under EHDS, datasets are expected to follow the FAIR principles, Findable, Accessible, Interoperable, and Reusable, through rich, machine-readable metadata and shared standards that enable seamless discovery and reuse. HealthDCAT-AP [4] is an application profile (a specialization) of DCAT-AP [2], and DCAT/DCAT-AP are defined as RDF vocabularies / RDF-based metadata standards [3]. HealthDCAT-AP standardizes how datasets are catalogued, including clear provenance and consistent semantics. A recent EHDS HealthDCAT-AP case study [5] shows how extending DCAT-AP with a minimal set of health-specific metadata can materially improve cross-border discoverability and interoperability in federated catalogues. Under EHDS Article 78, data owners should accompany each dataset with a data quality and utility label that explicitly includes a **bias examination** as part of data quality management, so users can judge representativeness and fitness for purpose. Bias examination is mandatory for publicly funded data sets. Metadata for each dataset are to be publicly available in an EU

Datasets catalogue. The QUANTUM Data Quality Labelling Tool [6] is designed to ensure that health datasets meet high quality standards, are readily accessible, and interoperate seamlessly across EU Member States. A key question that comes up when we consider bias examination is: "What makes a dataset diverse?" The Global Alliance for Genomics and Health (GA4GH) has published the Diversity in Datasets Policy v2.4 [7], arguing for the necessity of diverse datasets to address health equity. According to GA4GH, there is no single, universal definition of "diversity" in data; diversity is project specific. **Equity should be the central aim**: diversity is pursued to achieve, or at least move toward, more equitable health outcomes.

2. Methodology: Measuring Diversity of Health Datasets

We measure the diversity of a health dataset using the Shannon diversity index [8], which quantifies the entropy (or uncertainty) of a dataset. We compute it as:

$$H = -\sum_{\{i=1\}}^{S} p_i \ln p_i,$$

where p_1, p_2, ... , p_S are portions of different population groups with certain characteristics in the dataset. When only one group is present, $H = 0$, we have no diversity. The index is maximized when all groups are equal in size and grows with the number of groups. Suppose we are concerned with the diversity of age in the dataset. Then, the population groups are age quantiles: Group 1 is ages [1,25) and $p_1 = 0.2$ is the portion of people aged [1,25) in the dataset, group 2 is ages [25,50) with $p_2 = 0.5$, group 3 is ages [50,75) with $p_3 = 0.25$ and group 4 is ages [75,100) with $p_4 = 0.05$. Then, $H \approx 1.6$ bits. If all the groups were equally distributed, then we would have the maximum diversity at $H = 2$ bits. In the presence of multiple characteristics, the diversity index can be composite. For example, if we explored age and income, we would either request one index where the groups are selected according to both criteria, or a composite index with one value for age and another for income.

3. Results: Incorporating Diversity Metrics within a Dataset

We propose a two-step process for associating a diversity index with a health dataset. The first step is to include a metric of diversity along with the dataset's metadata in HealthDCAT-AP. This way, when a researcher searches the dataset catalogue for datasets to conduct their research on, they can get a first idea of how appropriate a dataset would be for their work. The second step includes a more fine-grained computation of a diversity index, specific to the researcher's purpose and involving information exchange between the data owner and the researcher using HL7 FHIR. HL7 FHIR is essential because this fine-grained diversity index must be computed from attributes, which are not typically present in dataset-level catalogue metadata. We can implement the calculation as a FHIR **Measure** and return only aggregated results via a **MeasureReport**, enabling data owners to compute the index locally.

3.1. Diversity in the Dataset Catalogue.

The EHDS dictates that datasets should be described in a dataset catalogue, using the HealthDCAT-AP metadata standard. At present, HealthDCAT-AP offers limited options for publishing diversity information. Some bias signals can be inferred from temporal and geographical metadata already included. However, we argue that for a more thorough study of diversity, bias information should be pre-computed and stored in the "publisher notes" attribute, accompanied by the research purpose. There are attributes that are very important to keep track of; we call these the **core diversity attributes**. These include **demographic attributes** (e.g. age, nationality, ethnicity, gender, sex, sexual orientation). Then, it's important to consider **socioeconomic factors** (e.g. education level, income, parental income, occupation). An important related index is the Deprivation Index: Launoy et al. [9] aimed to develop a "standardized measure" to tackle socioeconomic inequalities in cancer. Another important factor is **geography** (e.g. region, zip code, rural/urban). Finally, it's vital to consider **health status** (e.g. genetic factors, comorbidities, mental health, chronic diseases) and **environmental factors** (e.g. pollution, food culture, cultural background). In addition to the core attributes, it is crucial to recognize key inequalities and sources of bias on a case-by-case basis.

Next, we examine purpose-specific bias in two use cases, cardiovascular disease and pediatric kidney and liver cancer, and outline key factors that can elevate risk in each case. These lists are not exhaustive; rather, they underscore the importance of identifying and measuring biases and advancing research to better understand and address them.

Purpose-specific bias in cardiovascular disease. **Severe Mental Illness (SMI):** Patients with SMI, like schizophrenia, die on average 15–20 years earlier than the general population often due to cardiovascular disease [10], having twice the cardiovascular mortality rate of the general population [11]. **Sex:** Women have been historically underrepresented in cardiovascular research. In 740 cardiovascular trials from 2010-2017, and only 38.2% of the participants were women [12]. The problem starts with preclinical research focusing on male animals [13]. **Rheumatological Disease:** Patients with systemic lupus erythematosus and rheumatoid arthritis have 2x and 1.5x respectively, the risk of cardiovascular events [14]. **Polycystic ovary syndrome (PCOS):** Tay et al. [15] highlight the importance of recognizing PCOS as a significant risk factor for cardiovascular disease morbidity. **Experiencing violence:** Chandan et al. [16] show that women with a history of intimate partner violence have a 31% increased risk of subsequent cardiovascular disease. **Social isolation:** The American Heart Association [17] found a direct association between social isolation and coronary heart disease, highlighting the need for interventions to improve cardiovascular health for socially isolated individuals.

Purpose-specific bias in Pediatric Kidney and Liver Cancer. **Racial Disparity:** There are significant racial disparities for children with kidney cancer. These disparities appear in all stages of battling kidney cancer, from diagnosis, waitlisting for a transplant, transplantation, and follow up care. Harford and Laster [18] discuss such disparities. Charnaya et al. [19] reported that, within two years, Black children were 33% less likely and Hispanic children 18% less likely to be placed on the transplant waitlist. Laster et al. [20] reported that Black children had a 39% lower likelihood of receiving a transplant, and Hispanic children a 12% lower likelihood. Disparities are also found in preemptive transplantation, living vs diseased donor, and transplantation outcomes. **Socioeconomic Disparities:** Ebel et al. [21] found that children with liver cancer from socioeconomically deprived neighborhoods face higher risks of waitlist mortality, poorer medication adherence, and increased graft failure or post-transplant death. **Geography:** Adler et al. [22] showed that in liver transplantation, children that live >200 miles from a transplant center have a 75% increased risk of waitlist mortality. **Pollution:** Khan et al. [23] reported that parental pesticide exposure before conception or during pregnancy is associated with a higher risk of kidney cancer in their children.

3.2. Collaborative Diversity Index Computation using HL7 FHIR

It is important to enable collaboration between researchers and data owners to support the computation of a research-specific diversity index. The datasets should include information on core diversity attributes as well as research-specific factors known to elevate risk. Without access to these attributes and risk factors, it is not possible to carry out a sound bias examination tailored to the research purpose. To this end, in Figure 1, we show the workflow of our proposed computation. After the data owner publishes general diversity information on the dataset catalogue, which includes publicly accessible metadata (including a generic bias examination), a researcher can browse the catalogue and find the datasets most applicable to her research using HealthDCAT-AP. Then, to assess the diversity of the dataset specifically according to the groups she cares about, she can request a specific index from the data owner using HL7 FHIR. The proposed process builds trust accelerating use of health datasets for research. Note that this collaborative assessment of diversity does introduce privacy risks. Thus, depending on the local data policy, a data sharing agreement may be requested by the data owner to validate the research purpose and protect the data. The exchange between the data owner and the researcher is documented. Afterwards, the catalogue includes previous uses of the data without the diversity index since that may be used to derive personal health data.

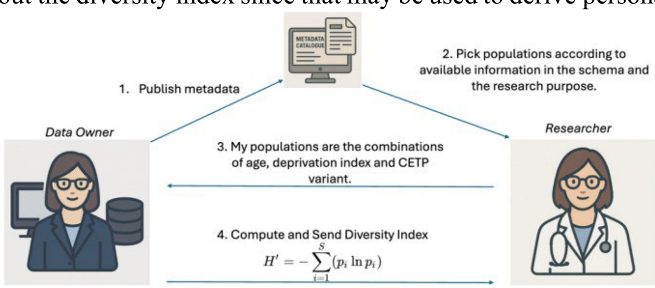

Figure 1: *Workflow:* **1.** Data owner *publishes dataset metadata*. **2.** Researcher *browses the metadata catalogue*. **3,4.** Researcher and data owner to c*ompute the diversity index according to research purpose.*

4. Conclusions

Diverse datasets are the backbone of excellent health research. The EHDS turns this principle into practice by pairing FAIR-aligned cataloguing in the EHDS with a quality-and-utility label that addresses bias assessment in health data sets. A pragmatic path forward is dual: publish high-level diversity signals in the catalogue for quick triage, then support purpose-specific, collaborative diversity assessments with researchers. With privacy protections in place, this approach accelerates discovery, helps derive robust models for applications in clinical practice, and raises the overall quality bar for health data. In addition to helping researchers select datasets appropriate for their research problem, the proposed approach operationalizes assessment of utility for health datasets. Making diversity measurable and publicly available, not only helps ensure research addresses underserved populations but also enables dynamic measurable updates to improve diversity properties of health datasets.

Acknowledgements

This work was supported by the European Union under the research projects (NextGen project GA:101136962), (Protect-Child GA101137423), (READi GA 101166227).

References

[1] European Council. European Health Data Space (2025). eur-lex.europa.eu/eli/reg/2025/327/oj/eng
[2] W3C (2024) Data Catalog Vocabulary (DCAT) - V3.2024 https://www.w3.org/TR/vocab-dcat-3/
[3] W3C (2013) RDF Schema 1.1 W3C Recommendation. 2013, https://www.w3.org/TR/rdf-schema/
[4] https://www.w3.org/TR/vocab-dcat-3/European Health Information Portal. HealthDCAT-AP (2024) healthinformationportal.eu/healthdcat-ap
[5] Derycke P, Barros BJ, Schutte NM, et al. (2025) Designing DCAT-AP extensions for common European data spaces: EHDS HealthDCAT-AP Case Study. InNeXt-generation Data Governance workshop 2025.
[6] QUANTUM: The health data quality label. https://quantumproject.eu/
[7] Global Alliance for Genomics and Health (2024) GA4GH Diversity in Datasets Policy v2.4.
[8] Shannon, C. E. (1948) A mathematical theory of communication. The Bell System Technical Journal
[9] Launoy G, Launay L, Dejardin O, et al. (2018). European Deprivation Index: designed to tackle socioeconomic inequalities in cancer in Europe. Eur J of Public Health. 2018 Nov 1;28(suppl_4)
[10] Polcwiartek C, O'Gallagher K, Friedman DJ, et al. (2024). Severe mental illness: cardiovascular risk assessment and management. European heart journal. 2024 Mar 21;45(12):987-97.
[11] Goldfarb M, De Hert M, Detraux J, et al. (2022). Severe mental illness and cardiovascular disease: JACC state-of-the-art review. Journal of the American College of Cardiology. 2022 Aug 30;80(9):918-33.
[12] Jin X, Chandramouli C, Allocco B, et al. (2020). Women's participation in cardiovascular clinical trials from 2010 to 2017. Circulation. 2020 Feb 18;141(7):540-8.
[13] Karp NA, Reavey N. (2019) Sex bias in preclinical research and an exploration of how to change the status quo. British journal of pharmacology. 2019 Nov;176(21):4107-18.
[14] Restivo V, Candiloro S, Daidone M, et al. (2022). Systematic review and meta-analysis of cardiovascular risk in rheumatological disease: Symptomatic and non-symptomatic events in rheumatoid arthritis and systemic lupus erythematosus. Autoimmunity reviews. 2022 Jan 1;21(1):102925.
[15] Tay CT, Mousa A, Vyas A, et al. (2023) international evidence-based polycystic ovary syndrome guideline update: insights from a systematic review and meta-analysis on elevated clinical cardiovascular disease in polycystic ovary syndrome. Journal of the American Heart Association. 2024;13(16):e033572.
[16] Chandan JS, Thomas T, Bradbury - Jones C, et al. (2020). Risk of cardiometabolic disease and all-cause mortality in female survivors of domestic abuse. J of the AHA. 18;9(4):e014580.
[17] Cené CW, Beckie TM, Sims M, et al. (2022). Effects of objective and perceived social isolation on cardiovascular and brain health: a scientific statement from the American Heart Association. Journal of the American Heart Association. 2022 Aug 16;11(16):e026493.Progress Made Toward Equitable Transplantation in Children and Young Adults with Kidney Disease
[18] Harford M, Laster M. (2024). Progress made toward equitable transplantation in children and young adults with kidney disease. Pediatric Nephrology. 2024 Sep;39(9):2593-600.
[19] Charnaya O, Zeiser L, Yisar D, et al. (2023). The unfinished journey toward transplant equity: an analysis of racial/ethnic disparities for children after the implementation of the Kidney Allocation System in 2014. Pediatric Nephrology. 2023 Apr;38(4).
[20] Laster M, Soohoo M, Hall C, et al. (2017). Racial–ethnic disparities in mortality and kidney transplant outcomes among pediatric dialysis patients. Pediatric Nephrology. 2017 Apr;32(4):685-95.
[21] Ebel NH, Lai JC, Bucuvalas JC, et al. (2022) A review of racial, socioeconomic, and geographic disparities in pediatric liver transplantation. Liver transplantation. 2022 Sep;28(9):1521-9.
[22] Adler JT, Bababekov YJ, Markmann JF, et al. (2017). Distance is associated with mortality on the waitlist in pediatric liver transplantation. Pediatric transplantation. 2017 Mar;21(2):e12842.
[23] Khan A, Feulefack J, Sergi CM. (2022). Exposure to pesticides and pediatric Wilms' tumor. A meta-analysis on pre-conception and pregnancy parental exposure with an IARC/WHO commentary. Human & Experimental Toxicology. 2022 Oct 17;41:09603271221136211.