JACC: ADVANCES VOL. 3, NO. 2, 2024

EDITORIAL COMMENT

Is Artificial Intelligence the Missing Link Between Administrative Data and Complete Patient-Level Health Records?*

Gerhard-Paul Diller, MD, PhD,[a,b] Ulrike M.M. Bauer, MD[b]

Previous studies have developed and utilized empirical algorithms based on clinician expertise to classify congenital heart disease (CHD) patients using claim and hospitalization data.[1-5] These algorithms have not only been tremendously clinically helpful but have also been validated and well-established in the literature. However, they are extremely time-consuming to implement, which naturally calls for an automated process. Marelli et al propose such a process, using machine learning (ML) methods to automate the coding process, thus making the identification of CHD patients more efficient.

In this issue of *JACC: Advances*, Marelli et al[6] offer a comprehensive investigation into the identification of CHD using sophisticated ML methods applied to large-volume administrative databases. The need for this innovative technology is grounded in the increasing reliance on administrative databases over the past decade for medical research and clinical care especially in the setting of CHD. In the absence of robust randomized controlled trials, these databases offer a wealth of information that can be leveraged to understand regional variations in diseases and support large-scale collaborative research initiatives. The study aims to utilize these databases to refine the identification of true CHD patients, which is crucial for improving patient care and outcomes. The authors employ a robust ML approach, using a cohort of approximately 20,000 patients from the established Quebec administrative database who have at least 1 CHD diagnosis. The cohort was randomly split into an 80% training set and a 20% test set. The training set was used to train the ML models, which included methods such as regularized logistic regression and decision trees. The models were evaluated with different training sample sizes to ensure sufficiency, and a 5-fold cross-validation method was used to establish optimal parameters. The area under the precision-recall curve was a key metric for model performance assessment.

The study demonstrates that the ML models can be sufficiently trained with sample sizes available in national CHD datasets. The models demonstrate good performance, with the Gradient Boosted Decision Trees model showing superior performance across several metrics. The Gradient Boosted Decision Trees model was especially noted for its ability to handle nonlinear data and capture high-order interactions among features, highlighting the potential of ML techniques over traditional parametric models. The study demonstrates the feasibility and the potential of ML methods to improve the accuracy of CHD diagnosis using administrative health databases, with implications for enhancing the standardization of electronic health record coding systems and the potential for generalization to other databases. Clearly, further research is needed to address the identified limitations and refine the ML models for broader applications, but the potential is more than evident.

ML algorithms have the capability to tackle complex analyses in the field of clinical science by using large datasets. These algorithms uncover patterns of interaction among various factors, allowing us to

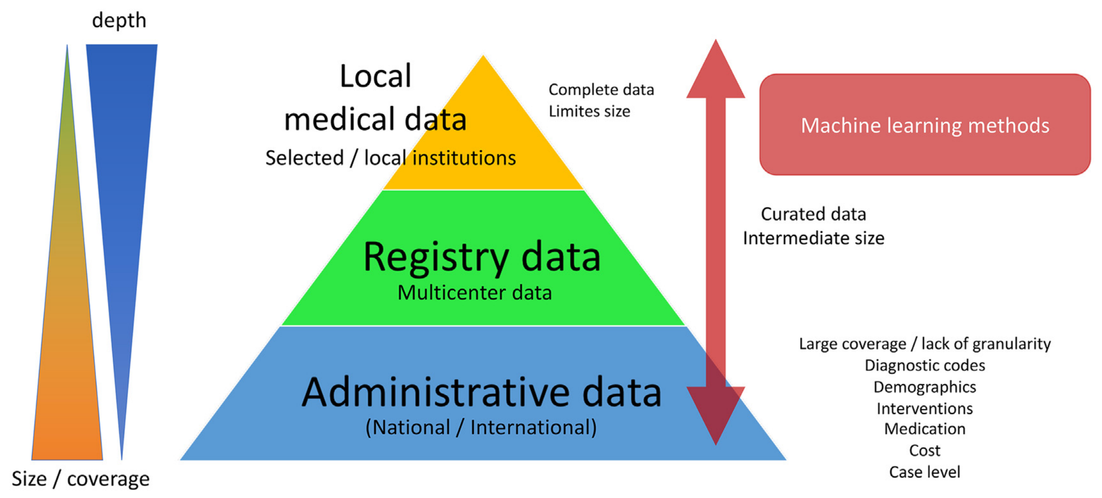
*Editorials published in *JACC: Advances* reflect the views of the authors and do not necessarily represent the views of *JACC: Advances* or the American College of Cardiology.

From the [a]Department of Cardiology III-Adult Congenital and Valvular Heart Disease, University Hospital Muenster, Muenster, Germany; and the [b]National Register for Congenital Heart Defects, Berlin, Germany.

The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

ISSN 2772-963X https://doi.org/10.1016/j.jacadv.2023.100802

FIGURE 1 **Potential Use of Machine Learning/Artificial Intelligence to Bridge the Size vs Granularity Depth in Congenital Heart Disease Research**



As a consequence, Marelli et al are to be commended for starting the process of big data integration in the setting of CHD. This approach is, in our view, the most likely source of relevant guidelines for clinical CHD studies in the years to come and an option to narrow the evidence gap between CHD and noncongenital cardiac disease. CHD = congenital heart disease.

surpass the limitations of cross-sectional risk prediction and progress toward the prediction of trajectories. Recent breakthroughs in deep learning (DL) neural networks have emerged as significant opportunities for acquiring knowledge from large datasets in the field of health care. While certain types of data in this context exhibit an organized and structured format, such as International Classification of Diseases-10 codes and imaging data, other types of data lack a structured format, such as clinical notes and electrocardiograms.

We have previously argued that, given the significance of longitudinal data and the extensive data available on the phenotypic characteristics of CHD, artificial intelligence (AI) may possess unique capabilities that make it highly suitable for analyzing this diverse population.[7] One must, however, also consider the tradeoff between data quantity and quality: Larger cohorts may exhibit less fidelity and detail compared to smaller cohorts, especially when administrative records are used. Advanced AI has the potential to improve this situation, as illustrated by the current report by Marelli et al.[6]

The next step will be to integrate natural language from electronic health record with administrative data. We have previously demonstrated the ability of DL models to utilize unprocessed data extracted from 44,000 raw medical records encompassing 10,019 patients diagnosed with CHD in a tertiary health care facility.[8] The primary objective of that study was to classify the patients into specific diagnostic groups, evaluate the complexity of their respective diseases, and determine their functional class according to the NYHA functional classification. Furthermore, the viability of employing DL techniques for the automated analysis of cardiovascular imaging investigations in CHD has been successfully showcased. The integration of these technologies has the potential to aid in effectively diagnosing CHD patients. Large CHD registries are required and combined with novel large language models will be at the center of implementing these emerging technologies to facilitate automated phenotyping of patients, hence enhancing the efficiency of longitudinal research.

As illustrated in Figure 1, AI has the potential to integrate various clinical data sources and provide efficient multimodal raw data analysis, including natural language processing to vertically integrate data sources and to add granularity to large-coverage administrative datasets.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

AI research at the National Register for Congenital Heart Disease and the University of Münster was supported by the Karla Völlm EMAH Foundation. The authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Prof Gerhard-Paul Diller, Department of Cardiology III-ACHD Unit, University Hospital Muenster, Albert Schweitzer Campus 1, A1, 48149 Muenster, Germany. E-mail: gerhard.diller@ukmuenster.de.

REFERENCES

1. Gilboa SM, Devine OJ, Kucik JE, et al. Congenital heart defects in the United States: estimating the magnitude of the affected population in 2010. *Circulation*. 2016;134:101–109. https://doi.org/10.1161/CIRCULATIONAHA.115.019307

2. Marelli AJ, Ionescu-Ittu R, Mackie AS, Guo L, Dendukuri N, Kaouache M. Lifetime prevalence of congenital heart disease in the general population from 2000 to 2010. *Circulation*. 2014;130:749–756. https://doi.org/10.1161/CIRCULATIONAHA.113.008396

3. Mylotte D, Pilote L, Ionescu-Ittu R, et al. Specialized adult congenital heart disease care: the impact of policy on mortality. *Circulation*. 2014;129:1804–1812. https://doi.org/10.1161/CIRCULATIONAHA.113.005817

4. Diller GP, Orwat S, Lammers AE, et al. Lack of specialist care is associated with increased morbidity and mortality in adult congenital heart disease: a population-based study. *Eur Heart J*. 2021;42:4241–4248. https://doi.org/10.1093/eurheartj/ehab422

5. Freisinger E, Gerss J, Makowski L, et al. Current use and safety of novel oral anticoagulants in adults with congenital heart disease: results of a nationwide analysis including more than 44 000 patients. *Eur Heart J*. 2020;41:4168–4177. https://doi.org/10.1093/eurheartj/ehaa844

6. Marelli AJ, Li C, Liu A, et al. Machine learning informed diagnosis for congenital heart disease in large claims data source. *JACC: Adv*. 2024;3:100801.

7. Diller GP, Arvanitaki A, Opotowsky AR, et al. Lifespan perspective on congenital heart disease research: JACC State-of-the-Art review. *J Am Coll Cardiol*. 2021;77:2219–2235. https://doi.org/10.1016/j.jacc.2021.03.012

8. Diller GP, Kempny A, Babu-Narayan SV, et al. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. *Eur Heart J*. 2019;40:1069–1077. https://doi.org/10.1093/eurheartj/ehy915

KEY WORDS administrative data, artificial intelligence, congenital heart disease, diagnosis