PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A Retrospective Comparison of Traditional and Artificial-Intelligence Based Heart Failure Phenotyping in a US Health System to Enable Real-World Evidence
AUTHORS	Garan, Arthur Reshad; Monda, Keri; Dent-Acosta, Ricardo; Riskin, Daniel; Gluckman, Ty

VERSION 1 – REVIEW

REVIEWER	Palmieri, Vittorio Ospedali dei Colli Monaldi Cotugno CTO, Cardiac surgery and transplantation
REVIEW RETURNED	21-Mar-2023

GENERAL COMMENTS	The matter of the study was to compare two different methods of evaluating the accuracy, i.e. the recall (or the sensitivity) and the precision (i.e. the positive predictive values), of the recognition of heart failure in the real world, to build a post-marketing real-world evidence. Structured and unstructured (narrative) data from electronic health records were used. Structured data were searched by looking at codes and a predefined list of problems, a modality that may be considered as a predefined fixed query system; unstructured data were searched by machine-learned, which resembles a multiparametric/probabilistic system to identify heart failure. Results reported in figure 3 are explicative: when it comes to search for specific key-words, single information (such as medications), the two methods are almost comparable in terms of accuracy. Artificial-based approach searching unstructured data was highly accurate in reaching the target of identifying heart failure subjects, the phenotype of heart failure (with reduced, preserved or mid-range ejection fraction), heart failure-related symptoms and medications. Traditional query modality was much less accurate, on the matter. What really limit the study, as per authors admission, is the lack of outcome data. Hence, impact and applicability of the results are essentially deduced. Data may incorporate some biases, as authors actually acknowledged. For instance, of the 472 cases identified, 81% where with reduced ejection fraction, which does not represent heart failure prevalence in population. Heart failure with preserved ejection fraction may be associated with events as frequent as heart failure with reduced ejection fraction. While I appreciated very much the study, I actuallay struggle to find it falling within the scope of the Journal.

REVIEWER	Sessa, Francesco

	University of Catania, Medical, Surgical Sciences and Advanced Technologies
REVIEW RETURNED	17-May-2023

GENERAL COMMENTS	This study aims to quantitatively evaluate the quality of data underlying real-world evidence (RWE) in heart failure (HF). The authors concluded that the use of an advanced, AI-based approach consistently identified HF phenotypes (i.e., HFrEF, HFmrEF, and HFpEF) more accurately than a traditional approach; moreover, common HF symptoms and comorbid conditions were consistently and accurately identified using an advanced approach; finally, medications for HF were accurately identified using both advanced and traditional approaches. The paper is interesting and well-written. In my opinion, it needs minor modifications before publication. The introduction section should be improved: this section should be functional to the study's aims. Moreover, I suggest improving the description of the aims. The material and methods section should be improved. I suggest a schematic picture to summarize the study's procedure, clarifying better the inclusion and exclusion criteria. The results section summarized the main findings. In the discussion section, the authors should improve the comparison of their data with the international data. As presented it is too redundant with the results section. Moreover, I suggest avoiding the use of the first person.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

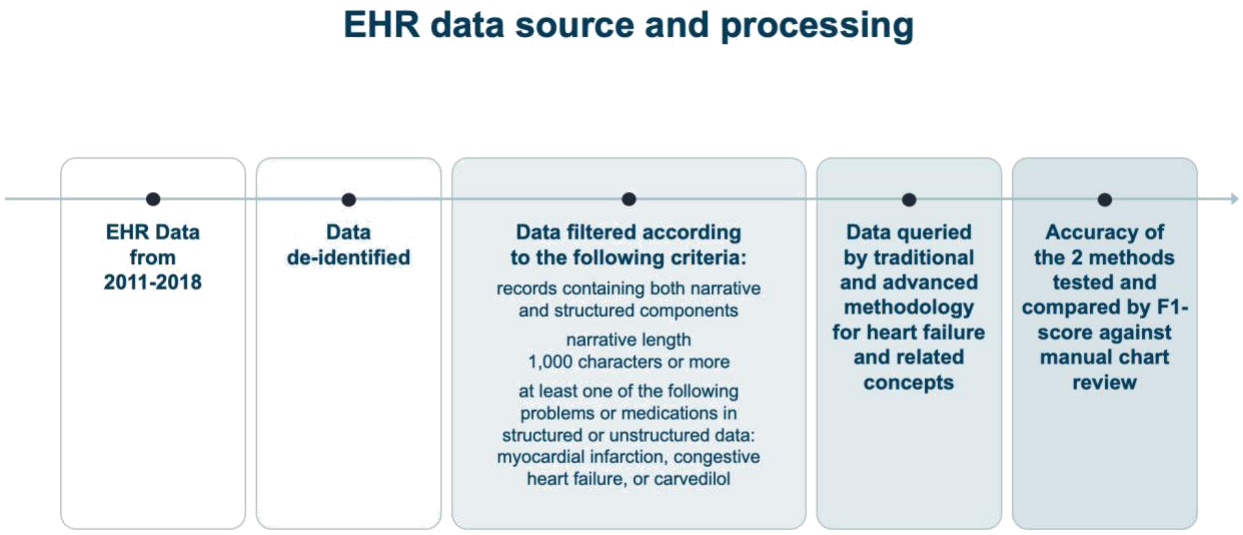
Comments to the Author:

The matter of the study was to compare two different methods of evaluating the accuracy, i.e. the recall (or the sensitivity) and the precision (i.e. the positive predictive values), of the recognition of heart failure in the real world, to build a post-marketing real-world evidence. Structured and unstructured (narrative) data from electronic health records were used. Structured data were searched by looking at codes and a predefined list of problems, a modality that may be considered as a predefined fixed query system; unstructured data were searched by machine-learned, which resembles a multiparametric/probabilistic system to identify heart failure. Results reported in figure 3 are explicative: when it comes to search for specific key-words, single information (such as medications), the two methods are almost comparable in terms of accuracy.

Artificial-based approach searching unstructured data was highly accurate in reaching the target of identifying heart failure subjects, the phenotype of heart failure (with reduced, preserved or mid-range ejection fraction), heart failure-related symptoms and medications. Traditional query modality was much less accurate, on the matter.

What really limit the study, as per authors admission, is the lack of outcome data. Hence, impact and applicability of the results are essentially deduced. Data may incorporate some biases, as authors actually acknowledged. For instance, of the 472 cases identified, 81% where with reduced ejection fraction, which does not represent heart failure prevalence in population. Heart failure with preserved ejection fraction may be associated with events as frequent as heart failure with reduced ejection fraction.

We appreciate the Reviewer's overview of our study and key findings. Indeed, the limitations we have cited are important ones for the reader to understand. First, to achieve a primary goal of understanding the accuracy of the data, manual chart abstraction was used. With a larger scale project, this would be a difficult task but we felt it a necessary part of the project as a whole before we analyze larger data sets across multiple centers. In order to facilitate the analysis to include manual chart abstraction, we sought to enrich the electronic health records included by restricting the cohort to those with certain criteria which we have outlined in the Methods section and added a new figure (included below) to make this more clear. For this reason, the population is skewed towards that of a heart failure with reduced ejection fraction cohort as the Reviewer notes.



Second, we do not include outcome data here, though this is not the primary focus of our analysis which was to test the accuracy of heart failure diagnoses derived from the electronic health record. As the Reviewer notes we have highlighted these limitations in our manuscript and explain that this is the first step in a multi-stage analysis.

While I appreciated very much the study, I actually struggle to find it falling within the scope of the Journal.

As real world evidence is increasingly used in all aspects of health care, not solely heart failure or cardiovascular disease, we sought to highlight the limitations of both traditional and AI-based methods of analyzing such data. Because regulatory bodies and insurance payors are increasingly utilizing real world evidence to make decisions regarding drug approval and reimbursement, we feel the limitations to the methodologies highlighted in this work are highly relevant to the journal's readership.

Reviewer: 2

This study aims to quantitatively evaluate the quality of data underlying real-world evidence (RWE) in heart failure (HF). The authors concluded that the use of an advanced, AI-based approach consistently identified HF phenotypes (i.e., HFrEF, HFmrEF, and HFpEF) more accurately than a traditional approach; moreover, common HF symptoms and comorbid conditions were consistently and accurately identified using an advanced approach; finally, medications for HF were accurately identified using both advanced and traditional approaches. The paper is interesting and well-written. In my opinion, it needs minor modifications before publication.

We are grateful for the Reviewer's appreciation for our work. We agree that the study is a relevant one particularly in light of increased utilization of real world evidence throughout all aspects of healthcare.

The introduction section should be improved: this section should be functional to the study's aims.

Moreover, I suggest improving the description of the aims.

We appreciate this suggestion. We have removed the following sentence to streamline the introduction:

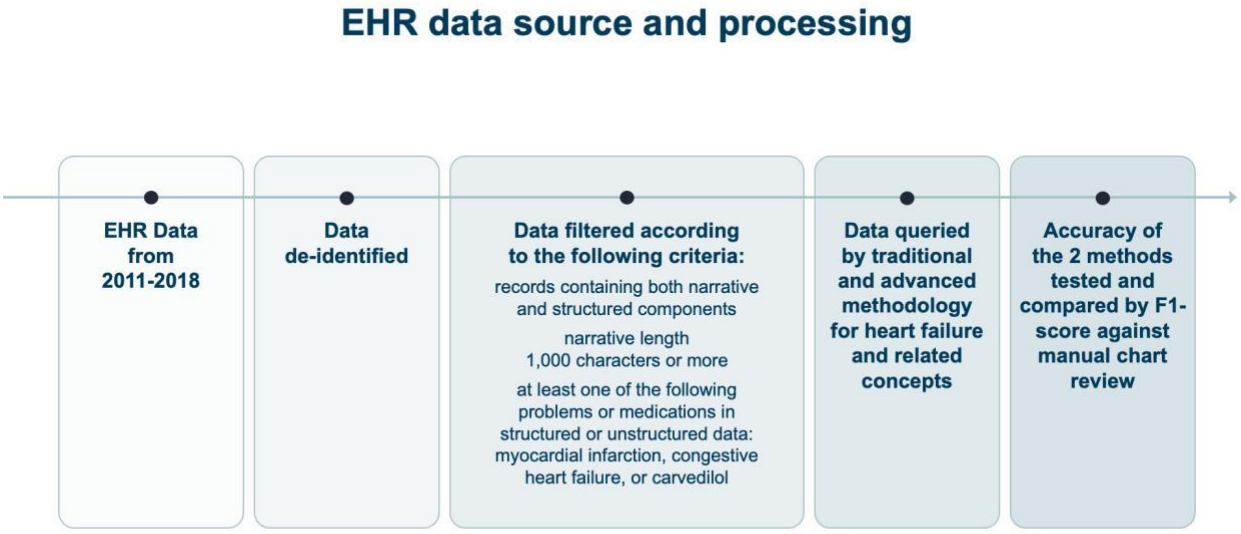
In contrast, registry data usually offers additional insights into more inclusive populations. Even with this, there is potential bias based on inclusion and exclusion criteria.

In addition, we have revised the final paragraph of the introduction as follows to better refine our primary aim of the study:

Traditional methods of identifying HF patients rely on querying diagnosis codes and structured data in the electronic health record (EHR) or medical claims. Conversely, artificial intelligence (AI) applied to unstructured data represents a novel method of analyzing the medical record. Because of the importance of data reliability in RWE and the potential to use unstructured data to achieve data enrichment[15], we sought to compare the accuracy achieved by traditional RWE methods versus advanced AI approaches in identifying a range of HF-specific data elements from the medical record.

The material and methods section should be improved. I suggest a schematic picture to summarize the study's procedure, clarifying better the inclusion and exclusion criteria.

We appreciate this suggestion from the Reviewer. We have added a Figure (Figure 1) to better describe the study methods as follows:



The results section summarized the main findings. In the discussion section, the authors should improve the comparison of their data with the international data. As presented it is too redundant with the results section. Moreover, I suggest avoiding the use of the first person.

We appreciate this suggestion from the reviewer and agree this is a valuable contribution to our manuscript. We have rewritten the Discussion, removing the use of the first person and have incorporated a comparison of international data. In particular, we have added the following seven references representing research from multiple nations including Canada, the United Kingdom, the Netherlands, and Sweden to strengthen our work.

Lim YMF, Molnar M, Vaartjes I, et al. Generalizability of randomized controlled trials in heart failure with reduced ejection fraction. Eur Heart J Qual Care Clin Outcomes. 2022;8(7):761-769.

Quach S, Blais C, Quan H. Administrative data have high variation in validity for recording heart failure. Can J Cardiol 2010;26:306–12. 10.1016/S0828-282X(10)70438-4.

Khand AU, Shaw M, Gemmel I, Cleland JG. Do discharge codes underestimate hospitalisation due to heart failure? Validation study of hospital discharge coding for heart failure. Eur J Heart Fail 2005;7:792–797.

Merry AH, Boer JM, Schouten LJ, Feskens EJ, Verschuren WM, et al. Validity of coronary heart diseases and heart failure based on hospital discharge and mortality data in the Netherlands using the cardiovascular registry Maastricht cohort study. Eur J Epidemiol 2009;24:237

Kao DP, Lewsey JD, Anand IS, et al. Characterization of subgroups of heart failure patients with preserved ejection fraction with possible implications for prognosis and treatment response. Eur J Heart Fail. 2015;17(9):925-35.

Uijl A, Savarese G, Vaartjes I, et al. Identification of distinct phenotypic clusters in heart failure with preserved ejection fraction. Eur J Heart Fail. 2021;23(6):973-982.

Ingelsson E, Arnlov J, Sundstrom J, Lind L (2005) The validity of a diagnosis of heart failure in a hospital discharge register. Eur J Heart Fail 2005;7:787–791.

VERSION 2 – REVIEW

REVIEWER	Palmieri, Vittorio Ospedali dei Colli Monaldi Cotugno CTO, Cardiac surgery and transplantation
REVIEW RETURNED	07-Jul-2023

GENERAL COMMENTS	I re-evaluated the study by A. Reshad Garan et al., and thank the Authors for the replies to criticisms and issues raised previously. I believe that the study is interesting. Yet, I found it not so oriented

	to clinical epidemiology or clinics of heart failure. Overall, no major changes have been made to the study. In particular, Authors were unable to provide outcome data. The explanation provided on the unbalanced proportion of HFrEF patients (81%) demonstrates that results of data searching, no matter the method used, are largely dependent on the quality of data, and query system. Difference in outcome of the two methods for searching phenotypes of HF si somehow a by-product of the two methodologies. I recognize that AI-based method to search for HF phenotypes has great potentials. Yet, data quality is an important determinant of the result of any query methodology, including AI-based, machine-learned-based, query procedure. I recognize the value of the work. My only question is the manuscript in its current version may be more suitable for Journals specialized in methodologies and results of Artificial Intelligence applied to Medicine.

REVIEWER	Sessa, Francesco University of Catania, Medical, Surgical Sciences and Advanced Technologies
REVIEW RETURNED	04-Jul-2023

GENERAL COMMENTS	The authors have improved their manuscript.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2

The authors have improved their manuscript.

We appreciate the reviewer's suggestions on the first review since these suggestions have allowed us to strengthen our work.

Reviewer: 1

I re-evaluated the study by A. Reshad Garan et al., and thank the Authors for the replies to criticisms and issues raised previously. I believe that the study is interesting. Yet, I found it not so oriented to clinical epidemiology or clinics of heart failure. Overall, no major changes have been made to the study. In particular, Authors were unable to provide outcome data. The explanation provided on the unbalanced proportion of HFrEF patients (81%) demonstrates that results of data searching, no matter the method used, are largely dependent on the quality of data, and query system. Difference in outcome of the two methods for searching phenotypes of HF si somehow a by-product of the two methodologies. I recognize that AI-based method to search for HF phenotypes has great potentials. Yet, data quality is an important determinant of the result of any query methodology, including AI-based, machine-learned-based, query procedure. I recognize the value of the work. My only question is the manuscript in its current version may be more suitable for Journals specialized in methodologies and results of Artificial Intelligence applied to Medicine.

We appreciate the Reviewer's perspective and appreciate the comment about the value of the work. The primary purpose of our analysis was to highlight the importance of data quality in generating evidence, rather than comparing patient outcomes. We respectfully disagree that data quality is "not so oriented to clinical epidemiology or clinics of heart failure" since data quality is at the core of clinical epidemiology. Regarding accuracy being influenced by the underlying data and query system, we wholeheartedly agree, which is the premise of the manuscript.

The topic is timely given recent retractions from NEJM and The Lancet related to data quality in real-world evidence. Lancet results were retracted on request of three authors who stated they could "no longer vouch for the veracity of the primary data sources." A NEJM publication was retracted because the authors were "unable to validate the primary data sources underlying our article."

With cardiovascular outcomes trials becoming increasingly expensive and life sciences firms increasingly turning to real-world evidence when required sample size is high, heart failure is seeing an increase in evidence generated from routine data. The credibility of underlying data and evidence is foundational for heart failure.

Since it may not be immediately apparent quite how much real-world evidence will influence the treatment of heart failure or how important data quality is, we have modified the following sentence in the discussion to highlight this point.

Furthermore, given the growth in RWE to support new drug indications, post-marketing surveillance, and decision-making regarding reimbursement, it is imperative for clinicians to understand that such inaccuracies may have a profound impact on large numbers of patients.