Identifying Heart Failure from Electronic Health Records: A Systematic Evidence Review

Rebecca T. Levinson[1], Jennifer R. Malinowski[2], Suzette J. Bielinski[3], Luke V. Rasmussen[4], Quinn S. Wells[5], Veronique L. Roger[6], Laura K. Wiley[7]

[1]Clinic for General Internal Medicine and Psychosomatics, Heidelberg University Hospital, Heidelberg, Germany
[2]Write InScite LLC, South Salem, NY, USA
[3]Division of Epidemiology, Mayo Clinic, Rochester, MN, USA
[4]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
[5]Division of Cardiovascular Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
[6]Division of Comprehensive Cardiology, Mayo Clinic, Rochester, MN, USA
[7]Division of Biomedical Informatics and Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Corresponding Author:
Laura Wiley, PhD
Assistant Professor
Division of Biomedical Informatics and Personalized Medicine
Department of Medicine
University of Colorado Anschutz Medical Campus
13001 East 17th Place, MS-F563
Aurora, CO 80045
Email: laura.wiley@cuanschutz.edu
Twitter: @GenomeGal

Funding: LKW was supported by the National Library of Medicine (NIH - K01LM013088) and institutional support from the Division of Biomedical Informatics and Personalized Medicine. LVR was supported by NHGRI grants U01HG008673-01 and NIGMS grant 2R01GM105688.

Acknowledgements: We would like to thank David Mayer and Melissa Wilson for providing project management support for this project.

ABSTRACT

Background: Heart failure (HF) is a complex syndrome associated with significant morbidity and healthcare costs. Electronic health records (EHRs) are widely used to identify patients with HF and other phenotypes. Despite widespread use of EHRs for phenotype algorithm development, it is unclear if the characteristics of identified populations mirror those of clinically observed patients and reflect the known spectrum of HF phenotypes.

Methods: We performed a subanalysis within a larger systematic evidence review to assess the different methods used for HF algorithm development and their application to research and clinical care. We queried PubMed for articles published up to November 2020. Out of 318 studies screened, 25 articles were included for primary analysis and 15 studies using only International Classification of Diseases (ICD) codes were evaluated for secondary analysis. Results are reported descriptively.

Results: HF algorithms were most often developed at academic medical centers and the V.A. One health system was responsible for 8 of 10 HF algorithm studies. HF and congestive HF were the most frequent phenotypes observed and less frequently, specific HF subtypes and acute HF. Diagnoses were the most common data type used to identify HF patients and echocardiography was the second most frequent. The majority of studies used rule-based methods to develop their algorithm. Few studies used regression or machine learning methods to identify HF patients. Validation of algorithms varied considerably: only 52.9% of HF and 44.4% of HF subtype algorithms were validated, but 75% of acute HF algorithms were. Demographics of any study population were reported in 68% of algorithm studies and 53% of ICD-only studies. Fewer than half reported demographics of their HF algorithm-identified population. Of those reporting, most identified majority male (>50%) populations, including both algorithms for HF with preserved ejection fraction.

Conclusion: There is significant heterogeneity in phenotyping methodologies used to develop HF algorithms using EHRs. Validation of algorithms is inconsistent but largely relies on manual review of patient records. The concentration of algorithm development at one or two sites may reduce potential generalizability of these algorithms to identify HF patients at non-academic medical centers and in populations from underrepresented regions. Differences between the reported demographics of algorithm-identified HF populations those expected based on HF epidemiology suggest that current algorithms do not reflect the full spectrum of HF patient populations.

INTRODUCTION

Heart failure (HF) is a complex syndrome where the heart is unable to fill with or eject blood sufficiently to meet the needs of the body. HF has a heterogeneous presentation, though dyspnea, fatigue, and fluid retention are common.[1] Many manifestations of HF exist, but HF subtypes are most commonly classified based on left ventricular ejection fraction (EF). HF subtypes differ in presentation and epidemiology: populations with HF with preserved EF (HFpEF) are commonly older, more female, and have more comorbidities than populations with HF with reduced EF (HFrEF).[1,2] HF patients experience significant morbidity and mortality, and both hospitalization and hospital readmission are common amongst HF.[3,4] In the US, it is estimated that medical care relevant to HF costs more than 30 million annually.[5] The combination of complex presentation and financial burden of HF has made understanding the disease in clinical populations a priority.

Although electronic health records (EHRs) are developed and maintained for patient care and clinical documentation, EHRs are increasingly finding a secondary use as a real world data source for studies of disease in clinical populations. HF is no exception with numerous studies developing algorithms to identify HF populations for the EHR[6,7] or using EHR identified HF populations for analyses across a variety of scientific domains.[8–12] Bolstered by targeted federal funding opportunities, there is significant interest in the growing field of HF phenomics to identify precise endophenotypes that capture the full spectrum of disease.[13,14]

Despite this renewed emphasis on deep phenotyping of clinically relevant HF populations, there has been little effort at systematically evaluating how the EHR-based phenotyping of HF has traditionally been performed and published. As such, it remains unclear whether populations identified by existing EHR algorithms represent the currently known spectrum of HF phenotypes and diagnosed patient characteristics. To elucidate this issue and provide a robust foundation for future deep phenotyping efforts in HF, we performed a systematic evidence review to evaluate how HF phenotypes are identified from the EHR and assess the utility of these algorithms for research and clinical application.

METHODS

Search Strategy

Search terms were refined with an iterative process involving co-authors RL, SB, LR, JM, and LW and incorporated Medical Subject Heading (MeSH) terminology to identify relevant secondary subject headings. This systematic evidence review (SER)was conducted as a subanalysis of a larger review investigating the quality of EHR-based cohort identification algorithms. Accordingly all studies were first captured by the search string used by the larger SER:

> ((((electronic health records OR "EHR" OR electronic medical records OR "EMR") AND ("natural language processing" OR "machine learning" OR classifier OR "deep learning" OR "artificial intelligence" OR phenotyp* OR "phenome" OR ICD OR probabilistic OR algorithm OR ("data-mining" OR "data mining")))))

Studies were further refined to only include those relevant to heart failure by applying the following search string:

> (((((("heart failure"[MeSH Major Topic]) OR ((cardiac or myocardial) AND (failure or insufficiency))) OR (congestive heart failure)) OR (congestive cardiomyopathy)) OR (cardiomyopath*)) OR ((cardi* or heart* or myocard*) AND (fail* or incompet* or insufficien* or decomp*)))

Search results for the larger project were uploaded to Covidence, a software for SER management. Heart failure search results were identified and prioritized for screening, review, and data extraction for this study.

Article Review

Search results were assessed in Covidence for inclusion in duplicate by blinded reviewers according to a pre-specified protocol. Conflicts were adjudicated through discussion or the input of a third reviewer, as necessary. Studies were included if they used EHR data to derive a phenotype that was a disease or condition, disease subtype, or disease symptom. Non-primary research articles, non-human studies, or those that were performed outside of the United States or those that only used claims data were excluded. Studies that did not describe identifying a patient population, identified the final patient population through International Classification of Disease Clinical Modification (ICD) diagnosis codes only or manual chart review, or simply applied previously published algorithms were also excluded. We reviewed all studies excluded for using a previously published algorithm to extract the algorithm citation and add to the overall study list if not already captured in our search strategy. For this subanalysis we excluded those studies that did not identify a heart failure phenotype, and we performed a secondary extraction of those studies excluded from the larger study that described extracting the HF population using only ICD diagnosis codes.

Data Extraction

As part of the larger study, reviewers (SB, RL, JM, LR, LW) extracted data in duplicate from included articles blinded to the other's responses using a two-stage process. First we extracted information about the focus of the article (i.e., algorithm method development, applied study, or portability of previously published algorithm method) and all eligible phenotype(s) within Covidence. We defined applied studies to be those where the primary purpose of the paper was to investigate a scientific question or describe a clinical application, rather than developing an algorithm to identify HF patients. Discordance between reviewers' responses were resolved through discussion or by author LW or JM as a third reviewer. Next, we extracted additional information about each study and algorithm into custom forms. Data extraction focused on identifying study details (e.g., study data source, type of applied study, etc.), algorithm methodology (e.g., type of data, type of computational method, etc.), algorithm validation (e.g., type of validation, process for performing manual review, etc.), and algorithm generalizability (e.g., demographics reporting, etc.). We extracted algorithm specific information individually for each qualifying algorithm. Neither quality assessment nor risk of bias were assessed for the included studies, as there is no validated instrument for these types of studies.

For this subanalysis we conducted additional data extraction for information specific to heart failure. For algorithm-based studies, we extracted who performed chart review validation, the performance characteristics of each algorithm, and demographic information (sex, age, race/ethnicity) for the population identified by the algorithm where available. To understand the relationship between algorithm-identified HF populations and those identified using only ICD diagnosis codes, we also performed a minimal data extraction on studies excluded in the full text review as ICD code only. This limited analysis only extracted the type of study, ICD-only phenotypes, whether the ICD definition was validated, and demographic information (sex, age, race/ethnicity) for the population identified by the algorithm where available. All subanalysis specific data extraction was performed via co-extraction by RL and LW.

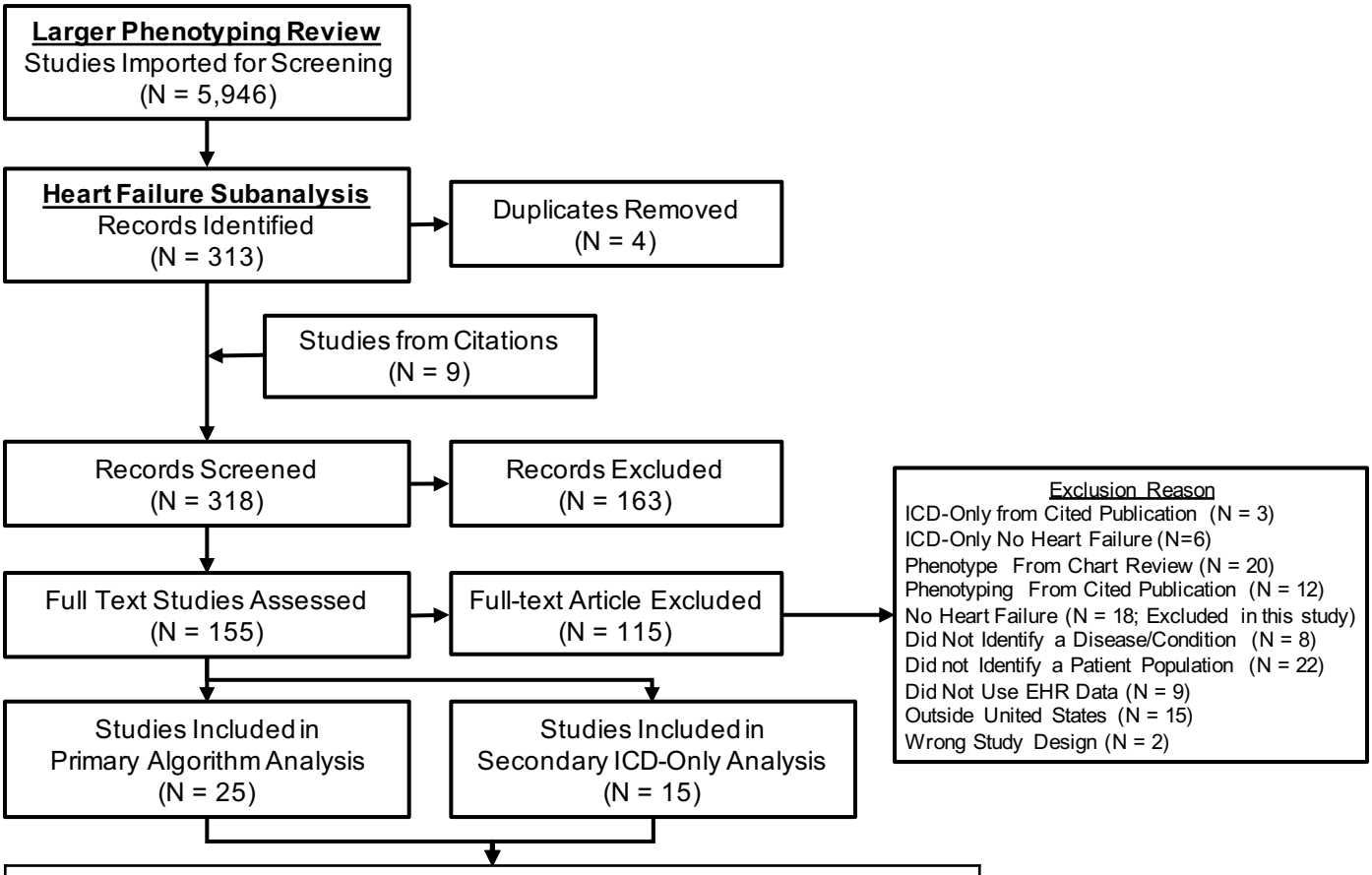
Analysis Plan

We analyzed and reported data extracted using descriptive statistics (counts, etc.) and visualization. We manually classified journals as "informatics" (e.g., focused on biomedical informatics or computer

science topics) or "clinical" (e.g., focused on clinical domains or healthcare delivery). We harmonized the EHR data source reported by each study (e.g., "Mayo Clinic EHR" and "Mayo Clinic hospital" harmonized to "Mayo Clinic"), classified the data source type (i.e., "Academic Medical Center", "Data Clearinghouse", "Health Information Exchange", "Regional Healthcare System", "Veterans Affairs", or "Other"), and identified the primary state where the EHR data source was located. For those studies reporting the definition used during manual review of patient records, two cardiologists (VR, QW) classified the provided definition by whether it met any standard diagnostic criteria. We performed all analyses using R version 3.6.0[15] and a variety of packages for data processing, graphing, and reporting.[16–24]

RESULTS

The initial literature search for the primary review was performed on January 29, 2019 and updated on November 13, 2020. Of the 5,946 studies assessed in the primary review, 313 were selected for this subanalysis using the query to identify heart failure specific studies. A further 9 studies that were identified from algorithm citations were added to the screening list. After removing duplicates, a total of 318 studies were screened for inclusion with 48.7% continuing to full text review. Full text review excluded 74.2% of the studies with the majority excluded for not identifying a patient population



Phenotype Identified	Algorithm	ICD Only
Heart Failure	10	13
Congestive Heart Failure	7	2
Heart Failure with Preserved Ejection Fraction	4	
Heart Failure with Reduced Ejection Fraction	3	
Acute Congestive Heart Failure	1	
Acute Decompensated Heart Failure	1	
Acute Heart Failure with Symptomatic Depressed Left Ventricular Contractility	1	
Borderline Heart Failure with Preserved Ejection Fraction	1	
Decompensated Heart Failure	1	
Heart Failure of Unknown Type	1	

Figure 1. PRISMA Diagram

(N=22) or identifying the phenotype using chart review (N = 20). **Figure 1** shows the PRISMA flow diagram providing a full accounting of all studies reviewed, excluded, and included in the final subanalysis.

After full text review, 25 studies were included in the primary algorithm-based analyses with a further 15 studies included in the ICD-only secondary analysis. From these 40 studies, 45 phenotype-study combinations were extracted from the combined study population, 30 identified with algorithms and 15 defined using only ICD codes. “Heart failure” and “congestive heart failure” were the most frequent phenotypes identified (17 with algorithms, 15 with ICD codes). Algorithm identified phenotypes also included those for identifying specific heart failure subtypes (N = 9) or acute heart failure (N = 4). **Figure 1** provides a complete list and frequency of phenotypes identified.

Heart Failure Detection Methods

To better understand how studies identified heart failure populations from EHRs we assessed the types of data and methods applied across the 30 algorithm identified phenotypes. The majority of

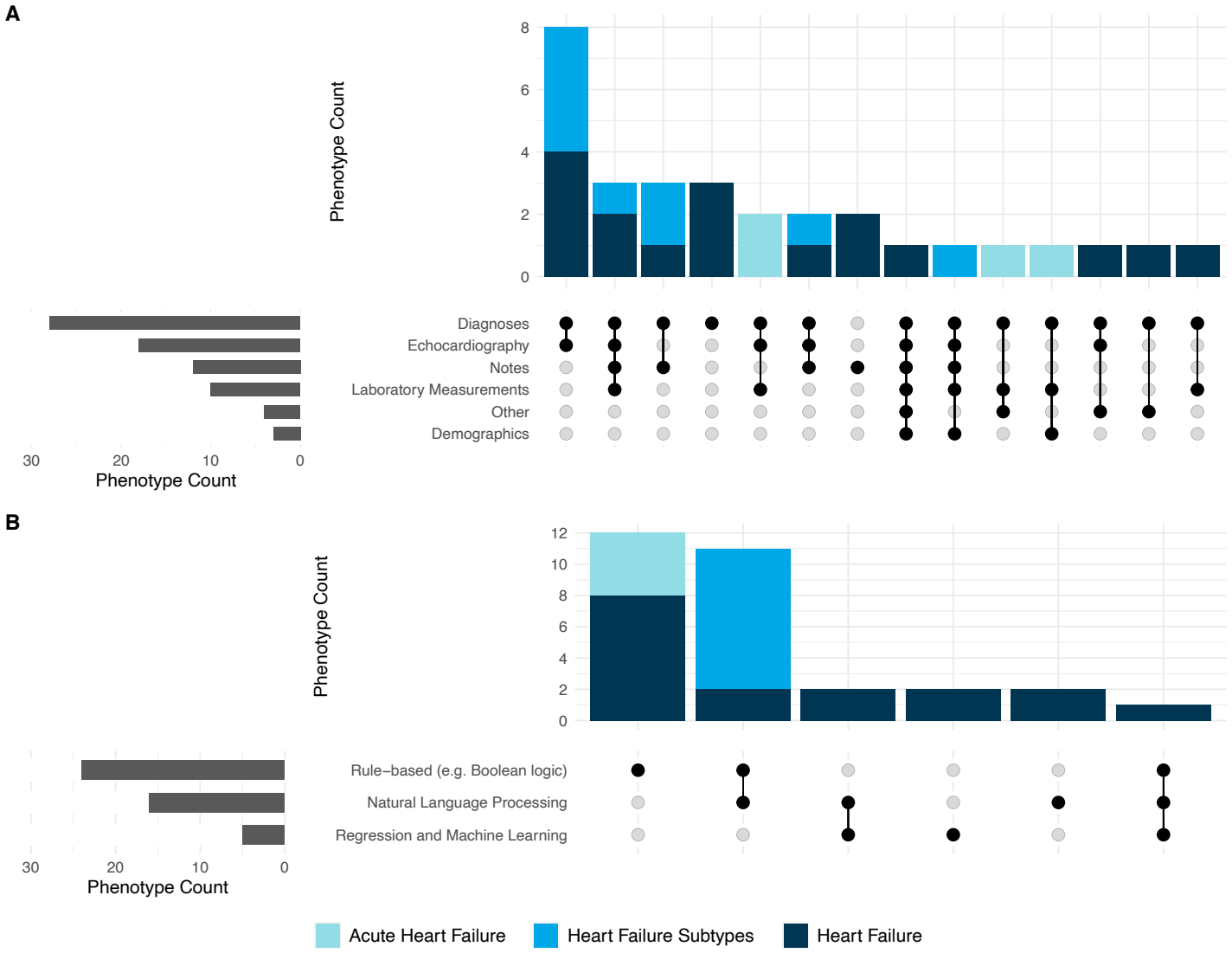


Figure 2. Algorithm Data Sources and Methods.
(A) Types of electronic health record (EHR) data used across the evaluated phenotypes (n=30). Black dots connected by lines show where multiple data sources are used (e.g., the first bar shows that both Diagnoses and Echocardiography were used); (B) Frequency of algorithm methods used across the evaluated phenotypes (n=30). Black dots connected by lines show where multiple methods were used (e.g., the second bar shows algorithms that used both Rule-based and Natural Language Processing methods).

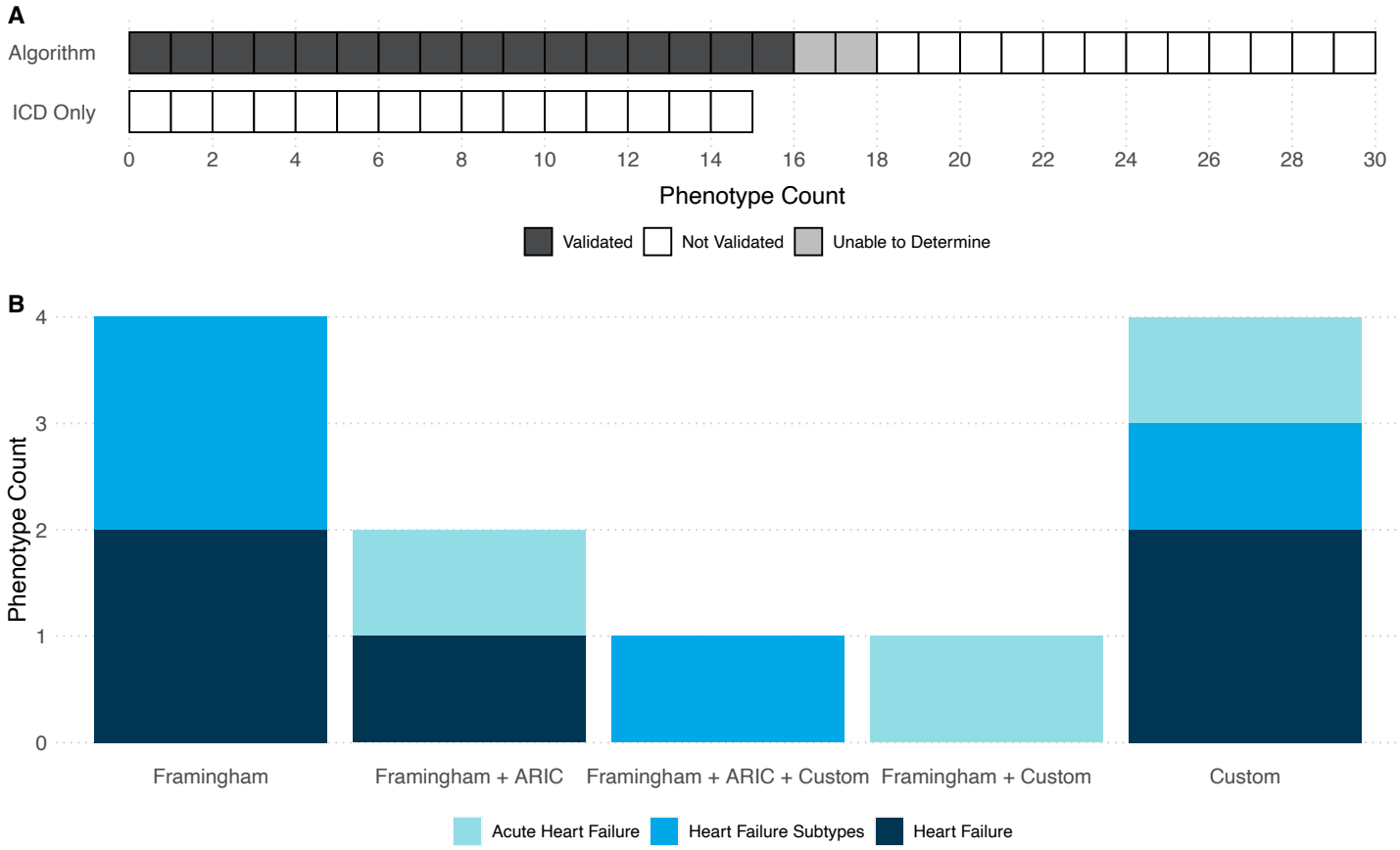


Figure 3. Heart Failure Definition and Validation.
(A) Number of phenotype algorithms that were validated. This includes the algorithm-based studies in our primary analysis (n=30), as well as the sub-analysis of studies that only used International Classification of Diseases (ICD) codes (n=15); (B) Frequency of review criteria used in the validation of phenotype algorithms, for the 12 phenotype algorithms that described their validation criteria. Studies may use a single criterion, or may have used a combination of criteria. Framingham = Framingham Heart Failure Diagnostic Criteria, ARIC = Atherosclerosis Risk in Communities, Custom = study-specific definition.

study phenotypes (N = 25) used two or more different types of data to identify HF patients. Diagnoses were the most common data type used across algorithms (N = 28) and most frequently used ICD codes (N = 24) or a combination of ICD and SNOMED codes (N = 4). Echocardiography data - either the presence of a report or extracted ejection fraction values - were the second most frequent data type (N = 18) and were used in algorithms detecting HF (50%, N = 9), HF subtype (77.8%, N = 7) and acute HF (50%, N = 2). **Figure 2A** reports the frequency of each data type and data type combinations used across study phenotypes.

The majority of study phenotypes used some combination of rule-based methods (N = 24) with 50% (N = 12) also using text extraction or natural language processing. A minority of study phenotypes were identified using regression or machine learning methods (N = 5). All acute HF study phenotypes were identified using exclusively rule-based methods, while all HF subtype study phenotypes were identified using a combination of text extraction and rule-based algorithms. **Figure 2B** reports the frequency of each methodology and methodology combinations used across study phenotypes.

Heart Failure Definition and Validation

To understand the potential applicability of each identified study phenotype for cardiovascular research, we assessed whether and how the ICD definition or algorithm were validated. **Figure 3A** reports the frequency of validation across all study phenotypes. Of the algorithm identified

populations, 52.9% of HF (N = 9), 44.4% of HF subtype (N = 4), and 75% of acute HF (N = 3) algorithms were validated. The majority (N = 15) validated algorithm performance using manual review of patient records, with 80% (N = 12) providing additional detail on how HF was defined during review. **Figure 3B** reports the frequency of formal diagnostic criteria used to define HF during review. The majority of phenotypes used Framingham HF diagnostic criteria alone (N = 4) or in combination with Atherosclerosis Risk in Community (ARIC, N = 2) or custom (N = 2) criteria.

Algorithm Development and Application

To better understand how HF algorithms and ICD-only definitions have been presented in the literature, we analysed study types and publication trends. **Figure 4A** reports the frequency of study type published over time separated by whether the study used an algorithm or an ICD-only definition. The studies identified in this analysis were fairly evenly split between algorithm method development (N = 18) and applied (N = 22). However within algorithm types, 64.0% of algorithm studies (N = 16) were method development focused, while 86.7% of ICD only studies (N = 13) were applied.

Next we analyzed the types of journals publishing these studies. The studies identified in this analysis were fairly evenly split between informatics (N = 21) and clinical journals (N = 19). However the majority of algorithm studies were published in clinical journals (52.0%, N = 13) while the majority of ICD only studies were published in informatics journals (60.0%, N = 9). **Figure 4B** reports the frequency of algorithm and study types presented in each type of journal. We further classified each applied study by all of the application domains (i.e., clinical evidence, clinical operations, epidemiology, prediction modeling, other, plus any combination). The most common applied study type was prediction modeling (N = 14). The most frequent combination of prediction modeling was applied to clinical operations (N = 5), with all of these studies using ICD only definitions of HF and publishing in informatics journals. **Figure 4C** reports the frequency of algorithm and applied study types presented in each type of journal.

Among HF algorithms we also considered whether the type of phenotype or validation status differed by study or publication type. The majority (58.8%, N = 10) of HF algorithms were published in informatics journals, while the majority of acute HF (75%, N = 3) and HF subtype (88.9%, N = 8) algorithms were published in clinical journals. Within informatics journals the two algorithms identifying acute HF and HF subtype were both part of method development papers. Within clinical journals the type of study (e.g., method development vs applied) was relatively similar within acute HF (N = 1 method development, N = 2 applied) and HF subtypes (N = 4 method development, N = 4 applied) algorithms. All HF subtype algorithms used in applied studies (N = 4) were used for a combination of clinical evidence and epidemiology, while all HF algorithms used in applied studies published in informatics journals (N = 4) were used for prediction modeling. **Figure 4D** reports the frequency of phenotype algorithm types by study and journal type. Among algorithms defined in method development papers, 100% of those published in clinical journals were validated (N = 10), compared to only 62.5% of those published in informatics journals (N = 5). Validation in applied studies was rare, with only a single instance of validation across 12 published algorithms. **Figure 4E** reports the frequency of algorithm validation by study and journal type.

Algorithm Generalizability

To better understand the potential generalizability of the HF algorithms identified we analyzed the locations and types of clinical data sources used to create each algorithm. The most common phenotypic data sources were academic medical centers (N = 13) and Veterans Affairs (VA) Hospitals (N = 5). **Figure 5A** reports the frequency of data source type for algorithm derived phenotypes. Although ten individual states were represented in the literature, the majority contributed to only a single algorithm. Minnesota was the most common site of algorithm development publishing

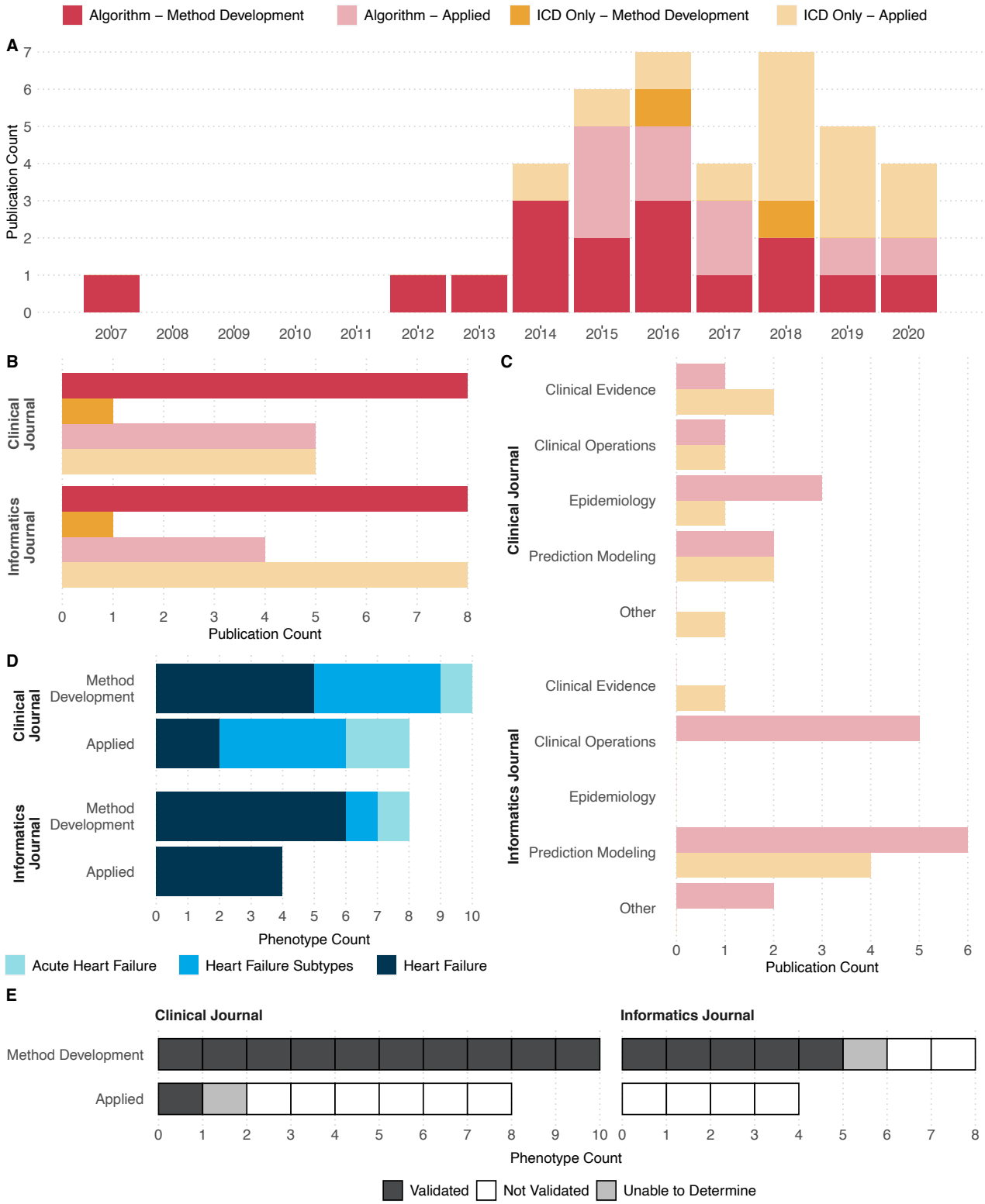


Figure 4. Algorithm Development and Application.

(A) Year of publication of analyzed studies, which includes the primary analysis set of phenotype algorithms (n=30) and the secondary analysis set of studies that used only International Classification of Diseases (ICD) codes (n=15). Studies are further subdivided if the primary focus was on algorithm method development or its application; (B) Frequency of studies published in a primarily clinical journal vs. a primarily informatics journal; (C) Frequency of the type of applied study published for algorithm or ICD-only phenotype definitions, subdivided by journal type; (D) Frequency of the study focus (method development vs. applied) across journal type (clinical vs. informatics), stratified by heart failure class for the primary phenotypes analyzed (n=30); (E) Frequency of phenotype definitions that were validated by journal type (clinical vs. informatics) and study focus (method development vs. applied).

10 algorithms, with the majority coming from the Mayo Clinic (N = 8). A total of 7 algorithms were developed with data from multiple states, the majority from the VA (N =5) and the remaining from data clearinghouses (N = 2). Additionally, 4 algorithms either did not list the EHR source at all (N = 3) or did not provide sufficient details about the EHR source to identify a location (N = 1). **Figure 5B** reports the frequency of data source location for algorithm derived phenotypes.

Understanding the population captured by an algorithm is essential to evaluating its generalizability. Therefore, we investigated which demographics were reported by each study. Demographics of any study population (not necessarily the identified HF population) were provided in only 68.0% of algorithm studies and 53.3% of ICD only studies. Regardless of HF identification approach, reporting study demographics was more common among studies published in clinical journals (Algorithm: 76.9%, N = 10; ICD Only: 66.7%, N = 4) compared to informatics journals (Algorithm: 58.3%, N = 7; ICD Only: 44.4%, N = 4). The minority of studies provided demographics for the identified HF populations, though reporting was more common among algorithm-defined phenotypes (40.0%, N = 12) compared to ICD only phenotypes (26.7%, N = 4). **Figure 5C** reports the frequency of reporting study and HF population demographics across algorithm and ICD only studies. For those HF populations with demographic reporting, 100% reported patient sex (N = 16), 81.2% reported frequency of at least one race or ethnicity (N = 13), and 75% reported mean or median age of the population (N = 12). The majority (N = 14) of HF definitions identified majority male populations (>50% male) including both algorithms detecting heart failure with preserved ejection fraction. **Figure 5D** reports the phenotype, data source, and reported demographics for available algorithm and ICD identified HF populations.

DISCUSSION

Heart failure is a syndrome diagnosed primarily through clinical assessment of patient symptoms and physical examination. Although multiple classification systems for HF exist with regards to functional status, left ventricular ejection fraction (EF) level, or clinical stage, there remains significant heterogeneity in the presentation, pathophysiology, and response to treatment.[1,25] Cluster analysis of clinical trial data has identified distinct subphenotypes that have different clinical outcomes, supporting the need for enhanced phenotyping of HF populations.[26,27] Because HF is a clinical diagnosis EHRs are a particularly valuable resource for deep phenotyping. While numerous studies have used EHRs to identify HF populations, the availability, utility, and quality of existing phenotyping algorithms are not well described. To provide a rational foundation for future phenomics projects,[13,14] we conducted a systematic evidence review of computational phenotyping in heart failure.

Our results demonstrate that while EHR-based phenotyping of HF has been an active area of research for many years, it remains an open area with multiple unexplored domains. HF phenomics dates back more than a decade, and has produced at least one published method development focused algorithm every year since 2012. There appears to be opportunity to identify more complex phenotypes as the majority of identified algorithms detected all-cause HF with a minority identifying specific HF subtypes or acute HF events. However even within all-cause HF, there remains significant heterogeneity in the types of data used across algorithms, suggesting that even broadscale population identification is not yet a settled task. Finally, although prediction modeling was a common application domain for HF populations, few of the identified studies actually used machine learning within their phenotyping algorithms. This was surprising as one would expect machine learning to be a useful method for capturing the complexity of HF. It is unclear whether this finding is due to limitations of our search strategy, a reflection of positive publication bias (e.g., ML methods have been applied, but performed poorly and were not published), or a true gap in the literature. Nevertheless, investigation of machine learning algorithms may present a rich opportunity for HF phenomics research.

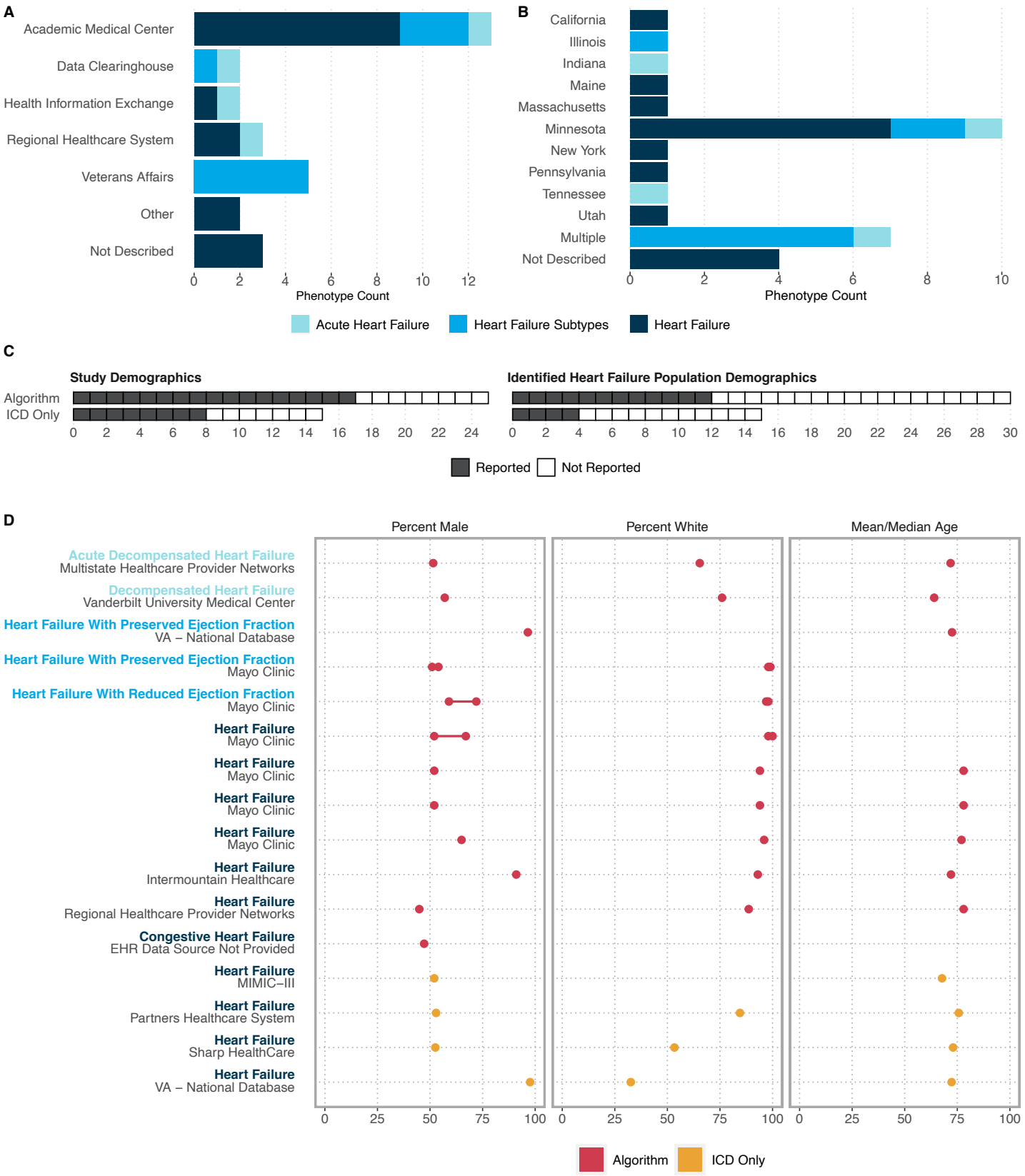


Figure 5. Study Population Characteristics.
(A) Frequency of EHR data source and (B) geographic location of the EHR data source population used to create each phenotype algorithm included in the primary analysis (n=30); (C) Frequency of the primary analysis algorithm definitions (n=30) and subanalysis International Classification of Diseases (ICD)-only phenotypes (n=15) that report any study demographics or demographics of the phenotype-identified heart failure population; (D) Phenotype description, data source, and reported demographics for both algorithm and ICD-defined phenotype definitions (when reported).

Another area of potential investigation is in the generalizability of HF algorithm performance across EHRs and utility for research and clinical applications. The second most common data type used across HF algorithms was the presence of, or EF measurement from, echocardiography reports. Although EF measurement is crucial to guide clinical treatment, patients treated at tertiary-referral centers may rely on outside reports that are rarely integrated into the clinical data warehouses potentially affecting performance of algorithms across institutions. Our ability to assess the utility of these algorithms for research and clinical application was hampered by the moderate rates of algorithm validation. Although one can use the logical combination of data types in algorithms to infer the patient population identified, validation serves an important role in tuning algorithm performance to the types of levels of evidence used during chart review to create a gold-standard case label. Just as clinical trial inclusion/exclusion criteria affect the generalizability of findings to the clinic, the phenotype definition used during validation determines whether an algorithm is fit for a particular analytic purpose. Positively, the majority of studies performing validation included the phenotype definition used during chart review. The most frequent diagnostic criteria used was Framingham either alone or in combination with ARIC and/or custom criteria. A significant minority of algorithms used definitions that did not correspond to standard diagnostic criteria. Interestingly, two used clinical trial definitions, which suggests the potential utility of these algorithms to demonstrate generalizability of trial findings in EHR populations. Investigators seeking to apply published algorithms should pay close attention to the phenotype definition used during validation to determine if the identified population is appropriate for the intended analysis.

In addition to the generalizability of the algorithm, we also considered whether the populations identified by published HF algorithms capture the full spectrum of HF patients. We see significant potential for selection bias given the reliance on echocardiography by the majority of algorithms. As described above, many patients at tertiary care facilities may only have outside test results available; this is especially concerning in light of the fact that the majority of sites developing phenotyping algorithms are academic medical centers with significant referral populations. It is possible that a number of HF patients are not being captured by these algorithms. Similarly, the concentration of algorithm development within a single state (Minnesota) and even a single institution (the Mayo Clinic) or in institutions with highly skewed populations (i.e., the VA) reduces the potential generalizability of these algorithms to community clinics and populations in underrepresented regions. Certainly these concerns were supported within the studies that reported algorithm identified populations. As expected, algorithms developed at institutions in low-diversity geographic regions identify majority white populations, while the VA identified nearly 100% male populations. However even after accounting for these expected deviations, there remains stark differences from demographics expectations, with only two studies identifying majority female populations despite the prevalence of HF in the United States being majority female (51.61%).[28] Large scale projects like the National Heart Lung and Blood Institute's HeartShare program should be particularly mindful of these concerns to ensure all HF patients benefit from their phenomics investment.

Our study also shines light on potential challenges in dissemination and implementation of HF phenomics. First, we were surprised to find that publication of method development studies was equal across clinical and informatics journals, with the majority of complex phenotypes published in clinical journals. Existing reviews of phenotyping methodology have focused solely on informatics and computer science journals,[29,30] an approach which our findings suggest may miss a number of complex algorithms. Similarly we expected more applied studies to be published in clinical journals, whereas our findings show the majority were published in informatics journals. We also noticed important trends in terms of the use of algorithms vs ICD only definitions across application domains. In areas where codes are more commonly used (e.g., claims data in epidemiology), or where HF was not the primary interest area (e.g., computer science development with an HF application) we saw more applied studies using ICD only HF definitions. However, we see a potential concern for

implementation of clinical evidence created using HF phenomics: although most clinical evidence studies used algorithms, the majority of clinical operations studies used ICD only definitions. It will be critical to ensure alignment between the patient populations used to develop treatment guidelines and those for whom EHRs can implement practice guidance.

This study also identified significant areas of opportunity for the field to improve the rigor and reproducibility of both algorithm development and the reporting of algorithm-based studies. Our investigative team with decades of combined experience in EHR-based phenotyping were unable to fully identify the data sources, methods used, or validation status for five algorithms. We were also surprised and concerned to find that even crucial study details like EHR data source or population demographics were often not reported. A third of algorithm-based studies and nearly half of ICD only studies reported no demographics for any population in the paper. While publishing the demographics of algorithm identified populations has not historically been expected, the lack of reporting limits the ability to determine the potential generalizability of the algorithms. Especially with a phenotype as complex as HF, demographics provide important insight into the specific subpopulations that an algorithm may be preferentially identifying.

A major strength of this study is the unbiased manner by which identified papers containing HF algorithms. However it is also possible that our search strategy did not identify all studies even if they contained algorithms that would have met criteria for inclusion. The lack of standard definitions and terminology for cohort identification[29] make complete ascertainment of phenotyping studies difficult. However, we used an interactive process with feedback and positive controls from the full author group to ensure relevant papers were included. Assessing the risk of bias or study quality, both for individual studies and for all studies across an outcome, is an essential component of SERs and provides context in qualitative and quantitative syntheses. Unfortunately no tool currently exists to measure the risk of bias for electronic phenotyping algorithm studies. In absence of such a framework, we are unable to assess the risk of bias or quality of the included literature beyond noting the poor reporting of basic study information (e.g., where the population is derived).

In conclusion, we have completed the first systematic evidence review of computational phenotyping in heart failure. Overall our data suggest that current EHR identified HF populations do not reflect the full spectrum of HF phenotypes or patient populations. We also found significant room for improvement in reporting of HF phenotyping algorithms and study populations. There appears to be significant opportunities to advance the field of HF phenomics in both phenotypic and algorithm complexity moving forward.

REFERENCES

1. Yancy Clyde W., Jessup Mariell, Bozkurt Biykem, Butler Javed, Casey Donald E., Drazner Mark H., Fonarow Gregg C., Geraci Stephen A., Horwich Tamara, Januzzi James L., Johnson Maryl R., Kasper Edward K., Levy Wayne C., Masoudi Frederick A., McBride Patrick E., McMurray John J.V., Mitchell Judith E., Peterson Pamela N., Riegel Barbara, Sam Flora, Stevenson Lynne W., Tang W.H. Wilson, Tsai Emily J., Wilkoff Bruce L. 2013 ACCF/AHA Guideline for the Management of Heart Failure. Circulation. American Heart Association; 2013 Oct 15;128(16):e240–e327.
2. Shah AM, Pfeffer MA. Heart failure with preserved ejection fraction: a forest of a variety of trees. European heart journal. 2014. p. 3410–3412. PMID: 24898553
3. Bergethon KE, Ju C, DeVore AD, Hardy NC, Fonarow GC, Yancy CW, Heidenreich PA, Bhatt DL, Peterson ED, Hernandez AF. Trends in 30-Day Readmission Rates for Patients Hospitalized With Heart Failure: Findings From the Get With The Guidelines-Heart Failure Registry. Circ Heart Fail [Internet]. 2016 Jun;9(6). Available from: http://dx.doi.org/10.1161/CIRCHEARTFAILURE.115.002594 PMCID: PMC4928632
4. Blecker S, Paul M, Taksler G, Ogedegbe G, Katz S. Heart failure–associated hospitalizations in the United States. J Am Coll Cardiol. 2013 Mar 26;61(12):1259–1267. PMCID: PMC3838728
5. Urbich M, Globe G, Pantiri K, Heisen M, Bennison C, Wirtz HS, Di Tanna GL. A Systematic Review of Medical Costs Associated with Heart Failure in the USA (2014-2020). Pharmacoeconomics. 2020 Nov;38(11):1219–1236. PMCID: PMC7546989
6. Blecker S, Katz SD, Horwitz LI, Kuperman G, Park H, Gold A, Sontag D. Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data. JAMA Cardiol. 2016 Dec 1;1(9):1014–1020. PMCID: PMC5289894
7. Tison GH, Chamberlain AM, Pletcher MJ, Dunlay SM, Weston SA, Killian JM, Olgin JE, Roger VL. Identifying heart failure using EMR-based algorithms. Int J Med Inform. 2018 Dec;120:1–7. PMCID: PMC6233734
8. Freiberg MS, Chang C-CH, Skanderson M, Patterson OV, DuVall SL, Brandt CA, So-Armah KA, Vasan RS, Oursler KA, Gottdiener J, Gottlieb S, Leaf D, Rodriguez-Barradas M, Tracy RP, Gibert CL, Rimland D, Bedimo RJ, Brown ST, Goetz MB, Warner A, Crothers K, Tindle HA, Alcorn C, Bachmann JM, Justice AC, Butt AA. Association Between HIV Infection and the Risk of Heart Failure With Reduced Ejection Fraction and Preserved Ejection Fraction in the Antiretroviral Therapy Era: Results From the Veterans Aging Cohort Study. JAMA Cardiol. 2017 May 1;2(5):536–546. PMCID: PMC5541383
9. Lagu T, Pekow PS, Shieh M-S, Stefan M, Pack QR, Kashef MA, Atreya AR, Valania G, Slawsky MT, Lindenauer PK. Validation and Comparison of Seven Mortality Prediction Models for Hospitalized Patients With Acute Decompensated Heart Failure. Circ Heart Fail [Internet]. 2016 Aug;9(8). Available from: http://dx.doi.org/10.1161/CIRCHEARTFAILURE.115.002912 PMCID: PMC4988343
10. Cox ZL, Lewis CM, Lai P, Lenihan DJ. Validation of an automated electronic algorithm and “dashboard” to identify and characterize decompensated heart failure admissions across a medical center. Am Heart J. 2017 Jan;183:40–48. PMID: 27979040
11. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. JAMA Netw Open. 2020 Jan 3;3(1):e1918962. PMCID: PMC6991258
12. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. Stud Health Technol Inform. 2015;216:40–44. PMCID: PMC4905764
13. RFA-HL-21-016: HeartShare: Next Generation Phenomics to Define Heart Failure Subtypes and Treatment Targets Data Translation Center (U54 Clinical Trial Not Allowed) [Internet]. [cited 2021 Jan 29]. Available from: https://grants.nih.gov/grants/guide/rfa-files/RFA-HL-21-016.html
14. RFA-HL-21-015: HeartShare: Next-Generation Phenomics to Define Heart Failure Subtypes and Treatment Targets Clinical Centers (U01 Clinical Trial Not Allowed) [Internet]. [cited 2021 Jan 29]. Available from: https://grants.nih.gov/grants/guide/rfa-files/RFA-HL-21-015.html
15. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available from: http://www.R-project.org/

w16. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. Welcome to the Tidyverse. JOSS. 2019 Nov 21;4(43):1686.
17. Allaire JJ, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R. rmarkdown: Dynamic Documents for R [Internet]. 2020. Available from: https://github.com/rstudio/rmarkdown
18. Kassambara A. ggpubr: "ggplot2" Based Publication Ready Plots [Internet]. 2020. Available from: https://CRAN.R-project.org/package=ggpubr
19. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: Visualization of Intersecting Sets, [Internet]. IEEE Transactions on Visualization and Computer Graphics. 2014. p. 1983–1992. Available from: https://doi.org/10.1109/TVCG.2014.2346248
20. Krassowski M. ComplexUpset [Internet]. 2020. Available from: https://doi.org/10.5281/zenodo.3700590
21. Bryan J. Googlesheets4: Access Google Sheets using the Sheets API V4 [Internet]. 2019. Available from: https://CRAN.R-project.org/package=googlesheets4
22. Firke S. janitor: Simple Tools for Examining and Cleaning Dirty Data [Internet]. 2021. Available from: https://CRAN.R-project.org/package=janitor
23. Wickham H. forcats: Tools for Working with Categorical Variables (Factors) [Internet]. 2020. Available from: https://CRAN.R-project.org/package=forcats
24. Müller K. here: A Simpler Way to Find Your Files [Internet]. 2020. Available from: https://CRAN.R-project.org/package=here
25. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, Falk V, González-Juanatey JR, Harjola V-P, Jankowska EA, Jessup M, Linde C, Nihoyannopoulos P, Parissis JT, Pieske B, Riley JP, Rosano GMC, Ruilope LM, Ruschitzka F, Rutten FH, van der Meer P, Authors/Task Force Members, Document Reviewers. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail. 2016 Aug;18(8):891–975. PMID: 27207191
26. Ahmad T, Pencina MJ, Schulte PJ, O'Brien E, Whellan DJ, Piña IL, Kitzman DW, Lee KL, O'Connor CM, Felker GM. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. J Am Coll Cardiol. 2014 Oct 28;64(17):1765–1774. PMCID: PMC4254424
27. Francis GS, Cogswell R, Thenappan T. The heterogeneity of heart failure: will enhanced phenotyping be necessary for future clinical trial success? Journal of the American College of Cardiology. 2014. p. 1775–1776. PMID: 25443697
28. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Shay CM, Spartano NL, Stokes A, Tirschwell DL, VanWagner LB, Tsao CW, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. Circulation. 2020 Mar 3;141(9):e139–e596. PMID: 31992061
29. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. United States; 2014;21(2):221–230. PMID: 24201027
30. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. 2018 Jul;1:53–68. PMCID: PMC6583807
31. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015 Jan 6;162(1):55–63. PMID: 25560714