Andreano *et al. Systematic Reviews* (2024) 13:313
https://doi.org/10.1186/s13643-024-02717-8

Systematic Reviews

RESEARCH Open Access

Diagnostic accuracy of case-identification algorithms for heart failure in the general population using routinely collected health data: a systematic review

Anita Andreano[1], Vito Lepore[2], Pietro Magnoni[1*], Alberto Milanese[1], Caterina Fanizza[2], Deborah Testa[1], Alessandro Musa[2], Adele Zanfino[1], Paola Rebora[3], Lucia Bisceglia[2] and Antonio Giampiero Russo[1] on behalf of the PROPHET-I study group

Abstract

Background Heart failure (HF), affecting 1–4% of adults in industrialized countries, is a major public health priority. Several algorithms based on administrative health data (HAD) have been developed to detect patients with HF in a timely and inexpensive manner, in order to perform real-world studies at the population level. However, their reported diagnostic accuracy is highly variable.

Objective To assess the diagnostic accuracy of validated HAD-based algorithms for detecting HF, compared to clinical diagnosis, and to investigate causes of heterogeneity.

Methods We included all diagnostic accuracy studies that utilized HAD for the diagnosis of congestive HF in the general adult population, using clinical examination or chart review as the reference standard. A systematic search of MEDLINE (1946–2023) and Embase (1947–2023) was conducted, without restrictions. The QUADAS-2 tool was employed to assess the risk of bias and concerns regarding applicability. Due to low-quality issues of the primary studies, associated with both the index test and the reference standard definition and conduct, and to the high level of clinical heterogeneity, a quantitative synthesis was not performed. Measures of diagnostic accuracy of the included algorithms were summarized narratively and presented graphically, by population subgroups.

Results We included 24 studies (161,524 patients) and extracted 36 algorithms. Algorithm selection was based on type of administrative data and DOR. Six studies (103,018 patients, 14 algorithms) were performed in the general outpatient population, with sensitivities ranging from 24.8 to 97.3% and specificities ranging from 35.6 to 99.5%. Eight studies (14,957 patients, 10 algorithms) included hospitalized patients with sensitivities ranging from 29.0 to 96.0% and specificities ranging from 65.8 to 99.2%. The remaining studies included subgroups of the general population or hospitalized patients with cardiologic conditions and were analyzed separately. Fourteen studies had one or more domains at high risk of bias, and there were concerns regarding applicability in 9 studies.

*Correspondence:
Pietro Magnoni
pmagnoni@ats-milano.it
Full list of author information is available at the end of the article

Discussion The considerable percentage of studies with a high risk of bias, together with the high clinical heterogeneity among different studies, did not allow to generate a pooled estimate of diagnostic accuracy for HAD-based algorithms to be used in an unselected general adult population.

Systematic review registration PROSPERO CRD42023487565

Keywords Diagnostic accuracy systematic review, Heart failure, Administrative health data, Health claims, Case-detection algorithms

Background

Heart failure (HF) was universally defined in 2021 as a "clinical syndrome with symptoms and/or signs caused by a structural and/or functional cardiac abnormality and corroborated by elevated natriuretic peptide levels and/or objective evidence of pulmonary or systemic congestion" [1] that may have diverse aetiologies. It is one of the most prevalent cardiac disorders, and as the population ages, its prevalence is increasing [2]. HF affects 1 to 4% of adults in industrialized countries, and among people aged 70 years and older, its prevalence climbs up to 10% or more [2, 3]. Estimates from Africa and South America are scarce [2]. To perform prevalence studies on large areas and real-world studies for newly introduced treatments, as well as to design and monitor health policy interventions, it is essential to detect patients with HF, rapidly and inexpensively, at the population level. Health administrative data (HAD) have become a popular tool for disease research and surveillance because they allow for timely, systematic, and cost-efficient population-level analyses [4, 5]. Hundreds of algorithms have been developed and are used to detect patients with a certain disease using HAD in different health systems [6–9]. These HAD-based case-identification algorithms typically involve a combination of billing claims, hospitalization records, outpatient specialist services, drug prescription data, and exemption from co-payments, linked at the individual level in either a deterministic or probabilistic way. The accuracy of such algorithms varies, as misclassification may occur, with differences related to several aspects [10]. First, there are characteristics of each particular disease, including how it can be accurately detected through billing codes at various levels of severity, the frequency of interactions with the healthcare system, and the level of care required. Second, the accuracy of detection depends on the quality and availability of several types of administrative data across different health systems and on their accessibility for the organization implementing the algorithm. Evidence suggests that linking multiple sources of information enhances sensitivity. Additionally, the characteristics of the population and the context of application (e.g., acute hospital vs. community care) influence the accuracy of HAD-based algorithms, due to variations in disease prevalence, severity, and the availability of diverse data sources.

HAD-based case-detection algorithms have also been developed for HF. Their reported accuracy varies greatly among studies, depending on the targeted HF stage (e.g., early, advanced, all), the included population (e.g., general vs. hospitalized), and the source of information used (e.g., hospital discharge records, outpatient databases, prescribed drugs). Particularly, sensitivity reported in published studies might range from 0 to over 90%, while specificity, when available, is consistently higher and often over 95% in studies performed in the acute hospital setting [11]. Moreover, a consistent proportion of these algorithms were developed and are used without having been appropriately validated using a clinical reference standard [12], implying that their true diagnostic accuracy is unknown. There are already a few published systematic reviews on this subject [11, 13–15]. However, they present some limitations, such as including non-validated algorithms or being limited to a particular country [12, 14] or a coding system [13, 15]. Additionally, summary estimates for various settings were pooled, and the causes of heterogeneity were not thoroughly investigated in those studies. A comprehensive and updated systematic review of validated case-detection algorithms, based only on HAD, analyzing algorithms developed for different settings separately and including an analysis of the causes of heterogeneity is therefore missing.

Objective

To determine the diagnostic accuracy of case-identification algorithms from administrative data, in comparison to clinical evaluation, for the detection of prevalent congestive HF in adult patients. Additionally, this study aims to evaluate the degree of clinical heterogeneity, and to identify and describe its underlying causes.

Materials and methods

This systematic review was registered in PROSPERO [16] (CRD42023487565) and written following the Preferred Reporting Items for a Systematic Reviews and

Meta-Analyses of Diagnostic Test accuracy Studies (PRISMA-DTA) guidelines.

Eligibility criteria

We did not apply any limitations to publication date, language, or publication status.

Types of studies

Because we anticipated that only a limited number of studies with one or more validated algorithms would be retrieved, we considered all diagnostic accuracy studies, including case–control studies, despite the well-established tendency of this design to overestimate diagnostic accuracy [17], and planned to perform a sensitivity analysis. We classified studies determining HF status using information acquired at a single time point as cross-sectional, while we defined as longitudinal studies that considered a lookback window and used all contacts with the health system within that time window to define the presence of HF. We excluded qualitative studies and studies that did not provide measures of accuracy for the proposed case-detection algorithm(s) in internal or external validation. In order to include the latest available evidence, even if not yet fully published, and to reduce possible publication bias, both full-length papers and proceedings from conferences were considered for the systematic review, provided that the other inclusion criteria were fulfilled [18].

Participants

Studies on the general adult population were included in the review. No age threshold was used to define adults, but we accepted the definitions of "adult population" given in the primary studies. Studies where participants were a subgroup of the general population (e.g., males only, patients with a specific chronic condition, hospitalized patients) were also included in the review. No other inclusion or exclusion criteria for the participants were applied.

Target condition

The definition of HF, before the 2021 statement [2], was heterogenous across countries. We consequently accepted the definitions of "HF" given in the primary studies and investigated this as a possible source of heterogeneity. We included in the review different stages of the disease and classified them according to definitions used in the primary studies into as follows: HF, incident HF, advanced HF, and HF with left ventricular systolic dysfunction (LVSD).

Index test

The index test was defined as a case-detection algorithm using routinely collected healthcare data (HAD) that may or may not involve data linkages across different data sources. Although it is sometimes difficult to define what HADs are with respect to electronic clinical data, in general, records pertaining to billing information or management of the health system are considered administrative records, while information derived from patient management is considered clinical [19]. In this review, algorithms were included only if they were completely based on data passively collected at any time in the hospital or from territorial care. Studies or algorithms based on clinical information obtained from electronic medical records (EMRs) or disease-specific registry data (e.g., cancer registries) were excluded. For example, algorithms including whether a diagnostic examination was performed (e.g., execution of echocardiography) were included, but algorithms including the results of the same diagnostic examination (e.g., ejection fraction value) were excluded. This choice was made in the interest of generalizability, as there are very few countries in the world collecting results of clinical examinations into electronic databases in a systematic and standardized fashion at the national level. On the opposite, databases containing summaries and/or billing codes of hospital discharge records, outpatient visits, and drug prescriptions are available in most countries, and they often adopt common coding systems, such as the International Classification of Diseases for diagnoses and procedures or the ATC classification for drug dispensations [20, 21].

Reference test

The reference standard was represented by clinical and/or instrumental diagnosis of the target condition (HF) performed by direct medical evaluation or review of paper/digital clinical records by any health professional or trained researcher. Studies using self-administered questionnaires and/or any form of patient-reported outcome measures only were not included.

Search strategy and study selection

The electronic database search was performed on 23 November 2023. The search strategy was designed to access both the PubMed and Embase full-text archives. It considered all the articles regarding HF detection algorithms from administrative data published until the day of the search, written in any language. The search string was developed by two of the authors (P. M., A. M.); examined, modified, and validated by a librarian; and then applied to the specified digital databases. The search string was divided into two parts to identify the following concepts: the first part aimed at retrieving studies developing and/or validating algorithms used for case finding and based on HAD as data sources, and the second part included terms aimed at identifying HF, the target condition. In addition to this search strategy, we manually checked the reference lists of all the studies examined as full text, after title and abstract screening, and the studies included in identified previous systematic reviews (Table 1). Additionally, we carried out a search of gray literature on the websites of selected national health quality agencies using the string “heart failure” [22–24].

The study selection process was performed using the software CADIMA [25]. Two review authors (P. M., A. M.) independently screened the titles and abstracts of the retrieved studies and discarded clearly irrelevant studies. Two different authors (A. A., V. L.) independently assessed the full texts of potentially relevant studies for inclusion and tracked reasons for exclusion. For the studies included in the review, the same two authors independently extracted data using a predefined data extraction form, structured with drop-down lists, after piloting on five studies. All discrepancies in judgement were resolved by discussion.

Data collection and data items

The following information was extracted for each study (see Additional file 1 for details): general study information, whether the described algorithm was presented and internally validated as an original algorithm within the same study or if the study performed the external validation of an algorithm developed elsewhere, and population characteristics (age, sex, specific subgroup). For each algorithm, we collected details on the index test (data source(s), coding system(s), time criteria/thresholds, algorithm); reference standard, including the referred guidelines and/or detailed criteria used to

Table 1 Search strategies

Embase	((‘case definition’:ab,ti OR ’case-definition’:ab,ti OR ’case detection’:ab,ti OR ’case-detection’:ab,ti OR ’case identification’:ab,ti OR ’case-identification’:ab,ti OR ’case-finding’:ab,ti OR ’case finding’:ab,ti OR ’case-ascertainment’:ab,ti OR ’case ascertainment’:ab,ti OR ’disease definition’:ab,ti OR ’disease-definition’:ab,ti OR ’disease detection’:ab,ti OR ’disease-detection’:ab,ti OR ’disease identification’:ab,ti OR ’disease-identification’:ab,ti OR ’disease-finding’:ab,ti OR ’disease finding’:ab,ti OR ’disease-ascertainment’:ab,ti OR ’disease ascertainment’:ab,ti OR ((‘detect*’:ab,ti OR ’defin*’:ab,ti OR ’identif*’:ab,ti OR ’find*’:ab,ti OR ’ascertain*’:ab,ti OR ’diagnos*’:ab,ti) AND ’algorithm*’:ab,ti)) AND (‘administrative claims (health care)’/exp OR ’administrative’:ab,ti OR ’claim*’:ab,ti OR ’routinely collected health data’/exp OR ’routinely collected’:ab,ti OR ’electronic data’:ab,ti OR ’computer* data’:ab,ti OR ’electronic health record’/exp OR ’electronic medical record system’/exp OR ’medical information system’/exp OR ’electronic health*’:ab,ti OR ’electronic medical’:ab,ti OR ’computer* health*’:ab,ti OR ’computer* medical’:ab,ti) AND (‘validation study’/exp OR ’reproducibility’/exp OR ’sensitivity and specificity’/exp OR ’predictive value’/exp OR ’receiver operating characteristic’/exp OR ’valid*’:ab,ti OR ’agree*’:ab,ti OR ’concordan*’:ab,ti OR ’reproducib*’:ab,ti OR ’sensitivity’:ab,ti OR ’specificity’:ab,ti OR ’accuracy’:ab,ti OR ’positive predictive value’:ab,ti OR ’ppv’:ab,ti OR ’negative predictive value’:ab,ti OR ’npv’:ab,ti OR ’diagnostic odds ratio’:ab,ti OR ’c index’:ab,ti OR ’c-index’:ab,ti OR ’auroc’:ab,ti OR ’roc curve*’:ab,ti OR ’roc analys*’:ab,ti OR ’receiver operating’:ab,ti OR ’receiver-operating’:ab,ti OR ’receiver operator’:ab,ti OR ’receiver-operator’:ab,ti OR ’area under curve’:ab,ti OR ’area under the curve’:ab,ti)) AND (‘heart failure’/exp OR ’heart failure’:ab,ti OR ’cardiac failure’:ab,ti OR ’ventricular failure’:ab,ti OR ’myocardial failure’:ab,ti OR ’heart insufficienc*’:ab,ti OR ’cardiac insufficienc*’:ab,ti OR ’ventricular insufficienc*’:ab,ti OR ’myocardial insufficienc*’:ab,ti OR ’heart dysfunction*’:ab,ti OR ’cardiac dysfunction*’:ab,ti OR ’ventricular dysfunction*’:ab,ti OR ’myocardial dysfunction*’:ab,ti OR ’heart disfunction*’:ab,ti OR ’cardiac disfunction*’:ab,ti OR ’ventricular disfunction*’:ab,ti OR ’myocardial disfunction*’:ab,ti OR ’congestive heart disease*’:ab,ti)
PubMed	(((case definition[Title/Abstract]) OR(case-definition[Title/Abstract]) OR(case detection[Title/Abstract]) OR(case-detection[Title/Abstract]) OR(case identification[Title/Abstract]) OR(case-identification[Title/Abstract]) OR(case-finding[Title/Abstract]) OR(case finding[Title/Abstract]) OR(case-ascertainment[Title/Abstract]) OR(case ascertainment[Title/Abstract]) OR(disease definition[Title/Abstract]) OR(disease-definition[Title/Abstract]) OR(disease detection[Title/Abstract]) OR(disease-detection[Title/Abstract]) OR(disease identification[Title/Abstract]) OR(disease-identification[Title/Abstract]) OR(disease-finding[Title/Abstract]) OR(disease finding[Title/Abstract]) OR(disease-ascertainment[Title/Abstract]) OR(disease ascertainment[Title/Abstract]))OR(((detect*[Title/Abstract]) OR(defin*[Title/Abstract]) OR(identif*[Title/Abstract]) OR(find*[Title/Abstract]) OR(ascertain*[Title/Abstract]) OR(diagnos*[Title/Abstract]))AND(algorithm*[Title/Abstract])))AND((Administrative Claims, Healthcare[MeSH Terms]) OR(administrative[Title/Abstract]) OR(claim*[Title/Abstract]) OR(Routinely Collected Health Data[MeSH Terms]) OR(routinely collected[Title/Abstract]) OR(electronic data[Title/Abstract]) OR(computer* data[Title/Abstract]) OR(Electronic Health Records[MeSH Terms]) OR(Medical Records Systems, Computerized[MeSH Terms]) OR(Health Information Systems[MeSH Terms]) OR(electronic health*[Title/Abstract]) OR(electronic medical[Title/Abstract]) OR(computer* health*[Title/Abstract]) OR(computer* medical[Title/Abstract]))AND((Validation?Studies as Topic[MeSH Terms]) OR (Reproducibility of Results[MeSH Terms]) OR (Sensitivity and Specificity[MeSH Terms]) OR (Predictive?Value?of Tests[MeSH Terms]) OR (ROC Curve[MeSH Terms]) OR (Area Under Curve[MeSH Terms]) OR (valid*[Title/Abstract]) OR (agree*[Title/Abstract]) OR(concordan*[Title/Abstract]) OR (reproducib*[Title/Abstract]) OR (sensitivity[Title/Abstract]) OR (specificity[Title/Abstract]) OR (accuracy[Title/Abstract]) OR (positive predictive value[Title/Abstract]) OR (PPV[Title/Abstract]) OR (negative predictive value[Title/Abstract]) OR (NPV[Title/Abstract]) OR (diagnostic odds ratio[Title/Abstract]) OR (c index[Title/Abstract]) OR (c-index[Title/Abstract]) OR (auroc[Title/Abstract]) OR (roc curve*[Title/Abstract]) OR (roc analys*[Title/Abstract]) OR (receiver operating[Title/Abstract]) OR (receiver-operating[Title/Abstract]) OR (receiver operator[Title/Abstract]) OR (receiver-operator[Title/Abstract]) OR (area under curve[Title/Abstract]) OR (area under the curve[Title/Abstract])) AND (((Heart Failure[MeSH Terms]) OR (heart failure[Title/Abstract]) OR (cardiac failure[Title/Abstract]) OR (ventricular failure[Title/Abstract]) OR (myocardial failure[Title/Abstract]) OR (heart insufficienc*[Title/Abstract]) OR (cardiac insufficienc*[Title/Abstract]) OR (ventricular insufficienc*[Title/Abstract]) OR (myocardial insufficienc*[Title/Abstract]) OR (heart dysfunction*[Title/Abstract]) OR (cardiac dysfunction*[Title/Abstract]) OR (ventricular dysfunction*[Title/Abstract]) OR (myocardial dysfunction*[Title/Abstract]) OR (heart disfunction*[Title/Abstract]) OR (cardiac disfunction*[Title/Abstract]) OR (ventricular disfunction*[Title/Abstract]) OR (myocardial disfunction*[Title/Abstract]) OR (congestive heart disease*[Title/Abstract])))

diagnose HF evaluation of the patient or his/her clinical records by a health professional); and available information on sample size, disease prevalence, counts of true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), and the following accuracy measures: sensitivity, specificity, positive and negative predictive values (PPV and NPV), F1 index, and diagnostic odds ratio (DOR). Ideally, we aimed at extracting from each primary study the algorithm, purely based on HAD, judged by the authors themselves as having the best diagnostic accuracy in terms of DOR. However, many of the primary studies did not make this judgment, and some of the studies reported many slightly different algorithms based on different data sources. Consequently, if more than one algorithm was reported, for each unique combination of data sources (e.g., hospital discharge record, hospital discharge record plus outpatient database), we extracted the algorithm with the highest DOR, with discussion between the two authors in case of disagreement. DOR was chosen as it is a widely used single indicator of diagnostic performance [26]. We chose to extract the best algorithm from each unique combination of data sources, as their diversity is a prominent factor affecting the relative threshold between sensitivity and specificity [10]. Following these criteria, the maximum number of extracted algorithms from a single study was 4.

Assessment of risk of bias and evidence quality

We assessed the risk of bias and applicability of each included primary study by using the Quality Assessment of Diagnostic Accuracy Studies v.2 (QUADAS-2) tool. The QUADAS-2 comprises four domains: participant selection, index test, reference standard, and flow and timing [27]. We tailored the QUADAS-2 to the review, as recommended [27], rephrasing three signalling questions: the two questions from the index test domain (1—"Was the algorithm applied without knowledge of the reference standard?" and 2— "Was code selection determined in advance?") and the second question from the reference standard domain (2— "Did not only patients presenting a certain diagnostic code received the reference standard?"; see Additional file 2). We also provided specific guidance on answering the signalling question "Were all patients included in the analysis?" from the flow and timing domain. We determined to answer "yes" to this question if all patients included in the sample used for validation (not necessarily overlapping with the whole study sample) were included in the analysis and "no" otherwise. Two review authors independently performed the evaluation, and disagreements were resolved by discussion.

The included studies were visually assessed for potential publication bias by Deeks' funnel plot, and the associated regression test of asymmetry was performed [28].

Diagnostic accuracy measures and data synthesis

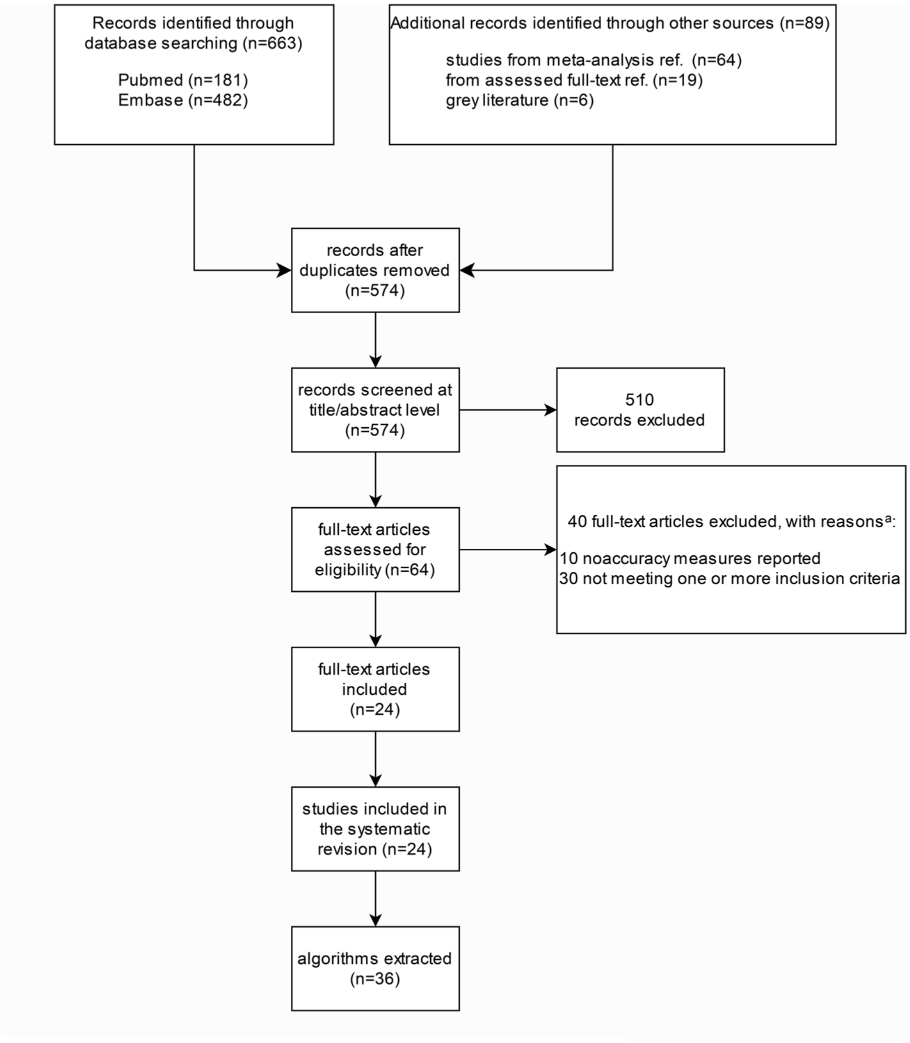
We evaluated the diagnostic accuracy through sensitivity and specificity. However, PPVs and NPVs were also extracted, or computed [29], and are reported in the summary tables in order to allow a complete evaluation and comparison with previous meta-analyses using PPVs. Additionally, PPV is the only diagnostic accuracy measure that can be computed from studies that only validated the conditions of subjects positive to the index test. These studies were reported and analyzed separately. We abstracted (or derived from tables) the numbers of TP, FP, TN, and FN from the full text, in order to reconstruct and verify the entire 2×2 table of diagnostic accuracy. We used those figures to calculate the measures of diagnostic accuracy that were not explicitly reported [29], including DOR, which was employed to choose which algorithms had to be extracted from each primary study. We contacted the corresponding authors via e-mail to obtain information needed to calculate the 2×2 table if missing or if incongruencies were detected. Confidence intervals for sensitivity, specificity, PPV, and NPV, if not reported, were calculated using the Wald method [30], which allowed to calculate the CI also when the entire 2×2 table could not be reconstructed. The study characteristics were summarized qualitatively. The extracted algorithms were presented by type of included patients: the general population (including the population recruited through territorial care or census registry), hospitalized patients, subgroups of the general population, and subgroups of hospitalized subjects. As described in the " Results" section, a high degree of both clinical and statistical heterogeneity, assessed through the I^2 statistic, was found in the included studies, also within population subgroups. More importantly, most of them were judged to be of low quality. Consequently, we decided not to perform the meta-analysis and to report ranges of sensitivity and specificity for the different population subgroups instead [31, 32].

We also presented the main results with a summary of findings (SoF) table, according to the Cochrane DTA Working Group approach, not using formal downgrading or overall credibility judgement [31].

Results

Search and study selection results

After duplicate removal, 574 potentially relevant studies were identified from all searched sources, 510 of which were subsequently excluded at the title and abstract


Records identified through database searching (n=663)
Pubmed (n=181)
Embase (n=482)
Additional records identified through other sources (n=89)
studies from meta-analysis ref. (n=64)
from assessed full-text ref. (n=19)
grey literature (n=6)
records after duplicates removed (n=574)
records screened at title/abstract level (n=574)
510 records excluded
full-text articles assessed for eligibility (n=64)
40 full-text articles excluded, with reasons[a]:
10 noaccuracy measures reported
30 not meeting one or more inclusion criteria
full-text articles included (n=24)
studies included in the systematic revision (n=24)
algorithms extracted (n=36)

[a] Details on reasons for exclusion are reported in Additional file 3

Fig. 1 Flow diagram of the study selection process based on PRISMA guidelines

screening stage (Fig. 1). Of the 64 full-text papers (including 5 conference abstracts or proceedings) assessed for eligibility, 40 were excluded for not fulfilling one or more inclusion criteria, as reported for each individual study in Additional file 3. Twenty-four studies were included in the systematic review, including 161,524 patients [19, 33–55]. None of the screened conference abstract or proceedings met the inclusion criteria; consequently, only full-length papers were included in the systematic review. Based on the criteria described in the " Materials and methods" section, after evaluating DOR within unique combinations of data sources, a total of 36 algorithms were extracted from the 24 included studies.

Characteristics of included studies

Table 2 outlines the characteristics of the 24 studies, published over three decades, from 1993 to 2022, and based on data from 1985 to 2018. Nine (38%) came from the USA, 8 (33%) from Canada, 6 (25%) from EU countries, and 1 (4%) from Australia.

Study design

There were 14 cross-sectional, 8 longitudinal, and 2 case–control studies.

Population included in the studies

Six studies (25%) included a general population recruited from primary or outpatient settings, and four additional studies (17%) were derived from the same population but were limited to subgroups: veterans (three studies, of which one included only hypertensive subjects) and patients with chronic obstructive pulmonary disease (COPD, one study). Eleven studies (46%) included hospitalized patients only, and the remaining 3 (13%)

Table 2 Characteristics of the included studies with respect to the sample used to assess diagnostic accuracy of the algorithms (i.e., internal or external validation)

Author, year	Country	Period	Study design	Population	*N* of pts	Age (mean/ median)	Females (%)	HF prevalence (%)	Type of reference standard[h] (criteria)
Dunlay, 2022 [33]	USA	2007–2017	Longitudinal	General	8657	74.1	49.8	9.8	RCR (ESC)
Vijh, 2021 [34]	Canada	2018–2018	Longitudinal	General, COPD	311	73.2	63.7	23.2	RCR (CCS, ESC, ACCF/AHA)
Xu, 2020 [35]	Canada	2015–2015	Cross-sectional[a]	Hospitalized	2105	64	50.2	14.1	RCR (ad hoc)
Cozzolino, 2019 [36]	Italy	2012–2014	Case control	Hospitalized	203	81.5c,e	53c	n.c	RCR (ESC)
Kaspar, 2018 [51]	Germany	2000–2015	Cross-sectional	Hospitalized	1042	77.5d,e	41d	21.3	RCR (cardiology expertise)
Tison, 2018 [37]	USA	2010–2012	Longitudinal	General	76,254	52.0e	52.9e	2.9	RCR (Framingham)
Franchini, 2018 [38]	Italy	2011–2014	Case control	General	389	76.3% > 65 yrs	48.1	n.c	RCR (cardiology expertise)
Bosco-Lévy, 2019 [50]	France	2014–2014	Cross-sectional	Hospitalized	229	69.1, 76.8f	28.8	n.c	RCR (ESC)
Blecker, 2016 [52]	USA	2013–2015	Cross-sectional[b]	Hospitalized	6549	60.9	50.8	NA	RCR (ARIC study)
Schultz, 2013 [39]	Canada	2004–2007	Longitudinal	General	2338	57.9	56	4.2	RCR (ad hoc)
Alqaisi, 2009 [40]	USA	2004–2005	Longitudinal	General	400	68e	54	65.0	RCR (Framingham)
Teng, 2008 [41]	Australia	1996–2006	Cross-sectional	Hospitalized[g]	1006	79.5e	49.7	n.c	RCR (Boston score)
Kümler, 2008 [53]	Denmark	1998–1999	Cross-sectional	Hospitalized	3201	70.8e	33.3e	13.4	Medical evaluation (ESC)
So, 2006 [42]	Canada	1994–2004	Cross-sectional	Hospitalized, AMI	193	68.0d,e	34.4d	28.5	RCR(*NR*)
Ingelsson, 2005 [43]	Sweden	1976–2001	Cross-sectional	Hospitalized	2322	NA	0	n.c	RCR (ESC)
Lee, 2005 [44]	Canada	1997–1999	Cross-sectional	Hospitalized	1641	75.5c,e	50.9c,e	n.c	RCR (Framingham)
Birman-Deych, 2005 [45]	USA	1998–1999	Cross-sectional	Hospitalized, AF	23,657	78.8	55	46.6	Pathology registry (*NR*)
Wilchesky, 2004 [46]	Canada	1995–1996	Longitudinal	General	14,980	NA	NA	7.1	RCR (primary care physician diagnosis)
Borzecki, 2004 [47]	USA	1998–1999	Longitudinal	General, veterans, HT	1176	NA	NA	7.0	RCR (attending physician diagnosis)
Quan, 2002 [54]	Canada	1996–1997	Cross-sectional[a]	Hospitalized	1200	NA	NA	10.7	RCR (Charlson)
Austin, 2002 [55]	Canada	1996–2000	Cross-sectional	Hospitalized	428	66.5	39	9.3	Pathology registry (attending physician diagnosis)
Szeto, 2002 [19]	USA	1996–1998	Longitudinal	General, veterans	148	64	4	10.1	RCR (*NR*)
Udris, 2001 [48]	USA	1996–2000	Cross-sectional	General, veterans	2246	68.8	3	34.6	RCR (ad hoc)
Jollis, 1993 [49]	USA	1985–1990	Cross-sectional	Hospitalized, CA	12,854	58.8	33.7	13.9	Research database (*NR*)

Abbreviations: ACCF/AHA American College of Cardiology Foundation/American Heart Association, *AF* atrial fibrillation, *AMI* acute myocardial infarction, *CA* undergoing coronary angiography, *CCS* Canadian Cardiovascular Society, *COPD* chronic obstructive pulmonary disease, *ESC* European Society of Cardiology, *HT* hypertension, *NA* not available, *n.c.* not meaningful to compute, *RCR* review of clinical records

[a] Quan H. et al. Med Care 2005; 43:1130. [b] Blecker S. et al. J Am Coll Cardiol. 2013; 61:1259. [c] Reported only for subjects positive to the algorithm. [d] Data relating only to subjects used for validation were not available; data for the whole population were extracted on the rationale that subjects used for validation were a random sample of the whole population. [e] Weighted mean of average in reported classes. [f] The first value is referred to the cohort used to evaluate Sn and Sp, the second one to a different cohort used to assess the PPV. [g] Nonelective admission only. [h] References for the criteria and ad hoc definitions of the reference standard are detailed in Additional file 4, Supplementary Table S1

included hospitalized patients with other cardiac conditions (acute myocardial infarction, $n=1$; atrial fibrillation, $n=1$; undergoing coronary angiography, $n=1$). Age varied from a reported mean or median of 52.0 years to 81.5 years, with two studies including people from the age of 40, one from the age of 50, and three from the age of 60–66 years. Four studies did not report any information on age. The percentage of females varied greatly, from 0 to 63.7%, with three studies not reporting information on sex composition of included subjects.

Target condition

HF was the target disease in 21 studies (88%). Incident HF, advanced HF, and HF with LVSD each were the target disease in a singular study.

Reference standard

Twenty studies (83%) reviewed clinical records to determine the presence of HF (reference standard). Of the remaining four studies, two were based on HF registries, one on medical evaluation, and one on a clinical research cardiology database. Details about the different definitions of HF adopted for the reference standard evaluation are reported in Supplementary Table S1 in Additional file 4. Only one study (4%) published in 2021 adopted the current globally accepted definition of HF [34]. Five studies (21%) referred to different versions of the ESC guidelines, three studies (13%) to the Framingham guidelines, three to other published and referenced definitions, three reported criteria used to define HF in the study without reference to a guideline, and nine (37%) stated that the diagnosis was established by a physician.

Index test

Hospital discharge data were included in 23 algorithms (64%) in 19 studies, always using codes from the ICD coding system, in its different versions (4 algorithms used only ICD-10, 3 used ICD-10 or −9, 15 used ICD-9, and 1 used ICD from 8 to 10). Eight algorithms (22%) in seven studies included outpatient diagnostic codes (one algorithm used ICD-10 or −9, and the rest used ICD-9 only). Drug prescription or dispensation databases were included in five algorithms (14%) in three studies, using either ATC or unspecified coding. Three algorithms (8%) in two studies included diagnostic and/or treatment information from a primary care database. One algorithm (3%) included emergency services data, and another one used exemption from co-payments. Fifty-five percent of algorithms ($n=20$) were based on a single data source, 28% on two types of data, and 17% on three. Further details on the developed algorithms and types of included data are reported in Supplementary Table S2 in Additional file 4.

HF prevalence

Excluding studies performed on subjects positive to the index test only (where the population prevalence cannot be assessed), reported HF prevalence in the population from which the validation sample was derived ranged from 2.9 to 65.0% overall. The median prevalence of HF in the validation sample was 5.6% (range 2.9–9.8%) in the 6 studies on general population and of 13.7% (9.3–100%) in the 8 studies including hospitalized patients.

Reporting of diagnostic accuracy measures

For the 6 studies (103,018 patients, 14 extracted algorithms) that were performed in the general outpatient population, the sensitivity ranged from 24.8 to 97.3%, and the specificity ranged from 35.6 to 99.5% (Table 3, Fig. 2). For the 8 studies including hospitalized patients and fully assessing diagnostic accuracy (14,957 patients, 10 algorithms extracted), the sensitivity ranged from 29.0 to 96.0%, and the specificity ranged from 65.8 to 99.2% (Table 4, Fig. 3). The 3 studies that only included a sample of subjects who were positive for the algorithm (2964 patients, 3 algorithms extracted) presented PPVs ranging from 82.0 to 99.5% (Table 4, Supplementary Fig. S1 in Additional file 4). The diagnostic accuracy measures of the remaining 7 studies (40,585 patients, 9 algorithms) are reported in Supplementary Table S3, Supplementary Fig. S2, and Supplementary Fig. S3 in Additional file 4. The studies including veterans ($n=2$) and hypertensive veterans ($n=1$) in the general population had a sensitivity ranging from 74 to 87% and a specificity ranging from 74.8 to 100%. The study with individuals with COPD ($n=1$) using data from a primary care database reported a sensitivity of 93.1% and a specificity of 90.8% for the best algorithm. The three studies including hospitalized patients with acute myocardial infarction ($n=1$), atrial fibrillation ($n=1$), or undergoing coronary angiography ($n=1$) reported sensitivities from 36.0 to 81.8% and specificities from 59.3 to 95.7%.

Assessment of bias

Study quality

Seventeen out of 24 studies (71%) were judged to be at high or unclear risk of bias in at least 1 domain. Patient selection was the domain with the highest percentage of studies at high risk of bias (eight studies, 33%), with one additional study having an unclear risk of bias (see Fig. 4 and Supplementary Table S4 in Additional file 4). Two studies did not include a random sample of patients: one randomly sampled patients but only among primary care physicians accepting to participate in the study [34], and the other chose to enrol patients presenting in a particular week [19]. Four studies applied inappropriate exclusions, discarding subgroups of subjects that would have

Table 3 Diagnostic accuracy results — general population

Study	Algorithm description in the primary study	TP	FP	FN	TN	N total	Sn	95% CI Sn		Sp	95% CI Sp		PPV	95% CI PPV		NPV	95% CI NPV	
Dunlay, 2022[a] [33]	D1 — 1 hospitalization for HF/VA in a year + 1 additional sign[b]	824	2421	23	5389	8657	97.3	97.0	97.6	69.0	68.0	70.0	25.4c	24.5	26.3	99.6c	99.5	99.8
	D2 — 1 hospitalization for HF/VA in a year or ED visit in HF/VA + 1 additional sign[b]	824	2851	23	4959	8657	97.3	97.0	97.6	63.5	62.5	64.5	22.4c	21.5	23.3	99.5c	99.5	99.7
	D3 — 2 hospitalizations for HF/VA in a year	628	836	219	6974	8657	74.1	73.2	75.0	89.3	88.6	90.0	42.9c	41.8	43.9	97.1c	96.7	97.4
	D4 — 2 hospitalizations for HF/VA in a year + 1 additional sign[b]	616	797	231	7013	8657	72.7	71.8	73.6	89.8	89.2	90.4	43.6c	42.6	44.6	96.9c	96.6	97.3
Tison, 2018 [37]	T1 — ≥ 1 HF code regardless of inpatient/outpatient status	1732	797	469	73,256	76,254	78.7	78.4	79.0	98.9	98.8	99.0	68.5	68.2	68.8	99.4	99.3	99.5
	T2 — ≥ 2 HF codes code regardless of inpatient/outpatient status	1358	354	843	73,699	76,254	61.7	61.4	62.0	99.5	99.4	99.6	79.3	79.0	79.6	98.9	98.8	99.0
	T3 — ≥ 1 inpatient or ≥ 2 outpatient HF codes	1613	564	588	73,489	76,254	73.3	73.0	73.6	99.2	99.1	99.3	74.1	73.8	74.4	99.2	99.1	99.3
Franchini, 2018 [38]	F1 — CARPEDIEM	271	50	41	27	389	86.7	83.3	90.1	35.6	30.4	40.8	84.5[c]	76.7	92.3	50.4c	45.4	55.4
Schultz, 2013 [39]	S1 — 1 CIHI or OHIP claim narrow	89	145	10	2094	2338	89.9	83.9	95.9	93.5	92.5	94.5	38.0	31.8	44.3	99.5	99.2	99.8
	S2 — 1 CIHI or 1 OHIP + 2nd claim (any source) in 1-year narrow	84	67	15	2172	2338	84.8	77.7	92.0	97.0	96.3	97.7	55.6	47.6	63.6	99.3	99.0	99.6
Alqaisi, 2009 [40]	A1 — ≥ 2 HF encounters OR any hospital discharge diagnosis of HF					100	70.8	61.9	79.7	74.3	65.7	82.9						
	A2 — ≥ 2 HF encounters OR primary hospital discharge diagnosis of HF					300	60.3	54.8	65.8	85.9	81.9	89.8						
Wilchesky, 2004 [46]	W1 — Diagnosis of HF, all physician	439	549	618	13,374	14,980	41.5	38.6	44.5	96.1	95.7	96.4	44.5	43.7	45.2	95.6	95.3	95.9
	W2 — Diagnosis of HF, billing physician	262	196	795	13,727	14,980	24.8	22.3	27.5	98.6	98.4	98.8	57.2	56.4	58.0	94.5	94.2	94.9

Abbreviations: *CIHI* Canadian Institute for Health Information, *OHIP* Ontario Health Insurance Plan, *VA* ventricular arrhythmia. Details on algorithms are reported in Supplementary Table S2 in Additional file 4

[a] Outcome, advanced HF. [b] Hyponatremia, hypotension, acute kidney injury/dialysis, use of high-dose loop diuretics, and use of metolazone. [c] Figures were recalculated using Sn and Sp values and HF prevalence reported in the study

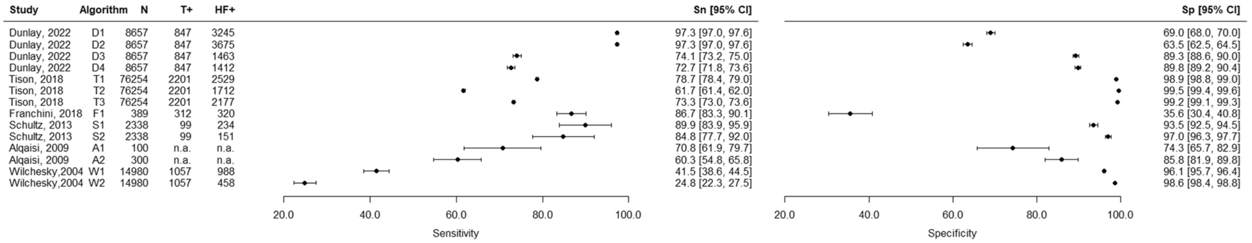


Fig. 2 Forest plots of sensitivity and specificity — studies on general population (see Table 3 for further details). N, total number; T+, algorithm positives; HF+, with heart failure

been more difficult to diagnose correctly [35, 40, 47, 48]. Finally, two studies employed a case–control design [36, 38]. Concerning the reference standard domain, four studies (17%) were affected by verification bias: in three studies, only index test-positive subjects were verified with the reference standard [41, 43, 44], allowing to estimate PPV only; in the fourth one, only subjects positive to a pre-screening received the reference standard assessment, thus artificially increasing prevalence [50]. Three studies (13%) also had an unclear risk of bias in this domain, as no enough information on how the reference standard was performed was present [42, 45, 49]. No risk of bias was detected concerning the index test domain. Issues were present in the flow and timing domain for seven studies (29%), where not all sampled patients received the reference standard [36, 37, 41, 43, 44, 49, 50, 50], mainly because the clinical records were not available for some patients. Also, in one of these studies, not all patients received the same reference standard [50], and in another one, not all patients were included in the final analysis [49].

Applicability

When evaluating applicability, patient selection was once again the most critical domain, with 38% of studies ($n = 9$) raising applicability concerns because of how and where subjects were recruited: in a second-level hospital [38], only nonelective hospital admissions [41], subjects affected by a particular comorbidity [34, 42, 45, 49], or veterans only [19, 47, 48]. Poor reporting issues were found, especially concerning patient selection. One study did not report the number of patients with the target disease.

Publication bias

In this systematic review, we included in the Deeks' funnel plot 27 out of 36 (75%) algorithms, for which it was possible to fully reconstruct the 2×2 table and calculate both the DOR and ESS. The funnel plot of these algorithms was substantially symmetric, and the regression test of asymmetry had a nonsignificant value (p-value = 0.99), indicating no evidence of a potential publication bias (Fig. 5).

Discussion

Summary of the main results

The results of our assessment of the diagnostic accuracy of validated HAD-based algorithms to detect HF, compared to clinical diagnosis, are summarized in the summary of finding (SoF) table (Table 5). The systematic review included 24 studies, with 161,524 participants. For the 36 HF case-detection algorithms analyzed, the reported range of sensitivity and specificity was from 24.8 to 97.3% and 35.6 to 100%, respectively. Summary estimates of sensitivity and specificity varied among population subgroups (Table 5). In all subgroups of patients and settings, we found a very high variability in both sensitivity and specificity.

Certainty of the evidence

This review was significantly limited by the inadequate reporting standards and general methodological weaknesses of several of the included studies. In our judgment, the risk of bias in the domains of patient selection and flow and timing was considerable, and there were applicability concerns regarding patient selection in a non-negligible number of studies. These studies were frequently conducted on samples with a much higher HF prevalence than either the general or hospitalized adult population. The reported biases were judged to hamper the validity of the summary estimates, especially due to the low numbers of studies in each population subgroup. First, the three studies verifying only subjects positive to the text would have been excluded from a meta-analysis [56]. Secondly, studies with poor quality of the reference standard, prone to non-differential misclassification bias, produce estimates of sensitivity and specificity that are lower compared to other studies [57]. As expected, case–control studies (also named two-gate design) inflated estimates of diagnostic accuracy compared with studies using a cohort of consecutive patients [58].

Combined with the clinical heterogeneity in both the index and reference tests, this means that it is not

Table 4 Diagnostic accuracy results — hospitalized patients

Study	Algorithm description in the primary study	TP	FP	FN	TN	*N* total	Sn	95% CI Sn		Sp	95% CI Sp		PPV	95% CI PPV		NPV	95% CI NPV	
Xu, 2020 [35]	X1 — ICD algorithm	170	14	126	1796	2106	57.4	51.8	63.0	99.2	98.8	99.6	92.4	88.6	96.2	93.4	92.3	94.5
	X2 — ICD algorithm in the nonsurgical patient cohort						60.0	54.2	65.8	99.1	98.6	99.6	93.2	89.5	96.9	92.1	90.7	93.5
Cozzolino, 2019 [36]	C1 — HF algorithm	119	8	5	71	203	96.0	91.0	99.0	90.0	81.0	96.0	94.0	88.0	97.0	93.0	85.0	98.0
Kaspar, 2018 [51]	K1 — ICD	110	7	112	813	1042	50.0	47.0	53.0	99.1	98.6	99.7	94.0	92.6	95.4	87.9	85.9	89.9
Bosco-Lévy, 2019 [50]	Diagnosis of HF in hospital discharge records	147		82		229	64.2	58.0	70.4									
		176	24			200							88.0	83.5	92.5			
Blecker, 2016 [52]	Bl1 — ICD9CM					315	71.4	66.4	76.4	98.7	97.4	100.0						
Kümler, 2008 [53]	Ku1 — Diagnosis of HF	126	30	303	2742	3201	29.0	27.4	30.6	99.0	98.7	99.3	81.0	79.6	82.4	90.0	89.0	91.0
Quan, 2002 [54]	Q1 — ICD-9	99	14	29	1058	1200	77.3	74.9	79.7	98.7	98.1	99.3	87.6	85.7	89.5	97.3	96.4	98.2
Austin, 2002 [55]	A1 — ICD-9 most responsible diagnosis	3204	1707	2271	51,634	58,816	58.5	58.1	58.9	96.8	96.7	96.9	65.2	64.8	65.6	95.8	95.6	95.9
	A2 — ICD-9 most responsible or secondary	4676	8375	799	44,966	58,816	85.4	85.1	85.7	84.3	84.0	84.6	35.8	35.4	36.2	98.3	98.1	98.4
PPV only																		
Teng, 2008 [41]	Diagnosis of HF	1001	5			1006							99.5	99.1	99.9			
Ingelsson, 2005 [43]	Diagnosis of HF	259	58			317							82.0	77.8	86.2			
Lee, 2005 [44]	ICD-9	1547	94			1641							94.3	93.2	95.4			

[a] Two different samples used to determine Sn (*n* total = 229, only with disease; not possible to compute Sp) and PPV (*n* total = 200, see note b). [b]Studies that selected a sample of patients positive to the algorithm and verified them to detect false positives only. Details on algorithms are reported in Supplementary Table S2 in Additional file 4

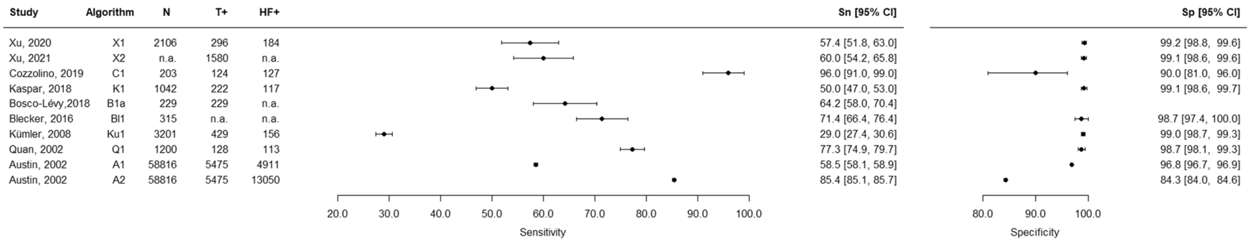


Fig. 3 Forest plots of sensitivity and specificity — studies including hospitalized patients (see Table 4 for further details). N, total number; T+, algorithm positives; HF+, with heart failure

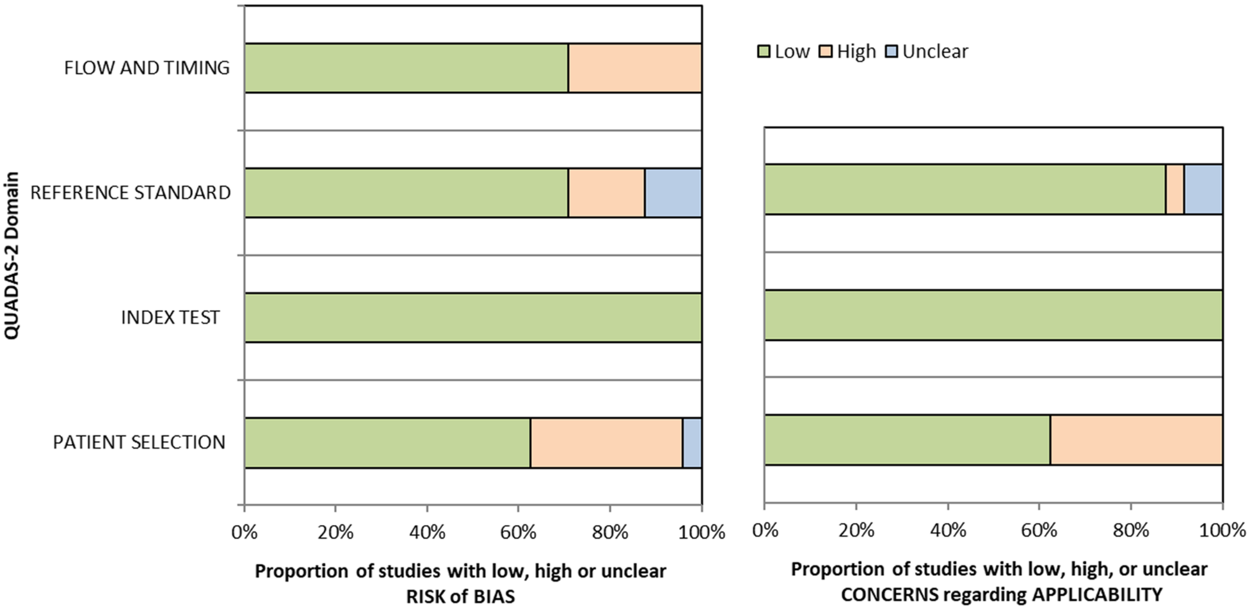


Fig. 4 Risk of bias (left panel) and applicability concerns (right panel) summary percentages across included studies, assessed and reported using the QUADAS-2 tool

possible to estimate the overall accuracy of HAD-based case-identification algorithms for HF in an unselected general adult population from the current literature. We were not able to assess the differential contribution of the different types of administrative data on the direction of sensitivity and specificity values, as very few studies ($n = 3$) reported diagnostic accuracy measures for different combinations of data sources.

Strengths and weaknesses of the review process

Although our database searches were reviewed and modified by a qualified librarian, approximately 90 of the screened articles were obtained from additional sources (Fig. 1). This limitation had already been found to a greater extent in a previous meta-analysis published in 2014 [11], which can be attributed to poor indexing of administrative data in digital databases until recently. In fact, the MeSH terms "Administrative Claims, Healthcare" and "Routinely Collected Health Data" were introduced only in 2016 and 2021, respectively. The studies retrieved manually were older than those retrieved from digital archives. An apparent limitation may be that we included only data derived from passively collected administrative data, excluding those algorithms that additionally employed the results of diagnostic tests to detect HF from EMRs, particularly B-type natriuretic peptide levels and ejection fraction percentages. However, even if these parameters are expected to improve the diagnostic accuracy of detection algorithms, EMRs both have privacy issues and lack standardization across hospitals or provider networks. These aspects limit their use, as they make access costly and time-consuming and reduce exportability.

Comparison with other studies and implications of the findings

Similar findings were reported in previous systematic review or meta-analyses. For example, Quach and

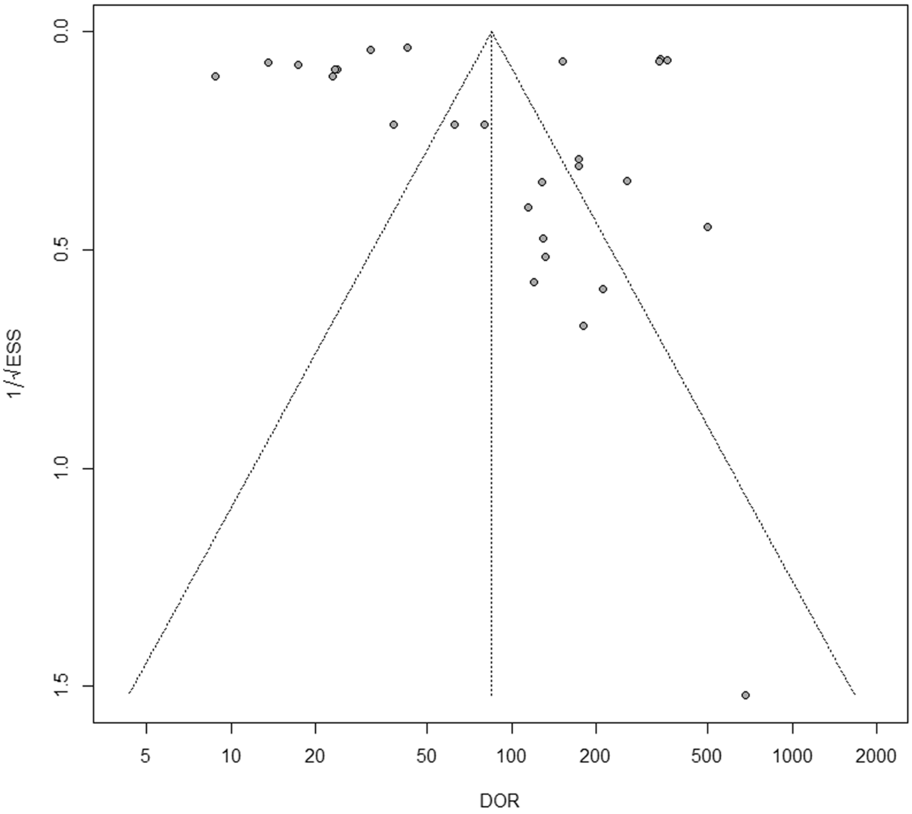


Fig. 5 Deeks' funnel plot to assess potential publication bias. The plot is substantially symmetric indicating no evidence of publication bias (regression test of asymmetry *p*-value = 0.99)

colleagues [13] investigated the accuracy of studies detecting HF with ICD codes in inpatient and outpatient claims up to 2008 and reported sensitivities ranging from 29 to 89%, with better specificities (always greater than 70% in studies based on hospital discharge records). McCormick and coauthors investigated studies up to November 2010 [11] and reported a pooled sensitivity of 75.3% (range 43–87%) and specificity of 96.8% (range 84–100%). Saczynski and colleagues [15] focused on PPV findings with more optimistic results (PPV from to 84 to 100%); however, their meta-analysis included many studies that used the reference standard to verify the disease status of only subjects testing positive to the algorithm (over 80% of studies reported PPV only) and also allowed acute HF diagnosis in the target disease definition.

Part of this heterogeneity is due to the intended use of the cohort of patients diagnosed with HF from HAD, which led to the development of algorithms that maximize either sensitivity or specificity. For example, if one wants to assess treatment effectiveness in a cohort of HF patients, they will develop an algorithm with high specificity at the expense of sensitivity. Conversely, if one wants to monitor HF prevalence over time, they will seek for a highly sensitive algorithm. Consequently, the intended use of the algorithm should be declared and pursued explicitly, as, for example, requiring more than one claim over a 1- or 2-year time frame will increase specificity but lower sensibility [37, 39], and the same may happen when adding drug prescriptions to the algorithm [34, 37, 48]. The discussed reviews [11, 13, 59] found, like ours, a high degree of heterogeneity in the types of HAD used in the algorithm, as well as in the definition of HF used in the clinical reference standard, the intended use of the algorithm, the setting, and the characteristics of the population. However, when analyzing studies that compared algorithms using data from multiple sources to those using only inpatient data, the former showed better overall diagnostic accuracy [33, 39, 40]. We performed a more rigorous systematic review compared to those previously published, avoiding methodological choices of potential concern such as including algorithms that were not validated, pooling together diagnostic accuracy estimates of algorithms developed and intended for use in different populations, and meta-analyzing PPVs of studies affected by verification bias together with those of studies correctly designed to estimate diagnostic accuracy. We also performed a broader systematic review, not

Table 5 Summary of findings table

Which is the diagnostic accuracy of validated algorithms based on health administrative data to diagnose heart failure compared to clinical diagnosis?					
Population		General adult population including subgroups based on demographics or comorbidities			
Prior testing		Some studies preselected subjects with a simpler version of the algorithm			
Setting		Primary care, outpatients or inpatients			
Index test		Case-detection algorithms from routinely collected health data			
Importance		Algorithms based on administrative health data are valuable to detect large cohort with heart failure rapidly and inexpensively			
Reference standard		Clinical diagnosis performed by a clinician or health professional (medical examination or medical chart review)			
Studies		14 cross-sectional studies, 8 longitudinal studies (using multiples contacts with the health system over 1–3 years), and 2 case–control studies			
Subgroup	Accuracy (95% CI)	No. of participants (studies, algorithms)	Prevalence in the sample used for validation of the algorithm Median (range)	Practical implication	Quality and comments
Outpatient or primary care general population	Range • Sensitivity 24.8% (95% *CI* 22.3–27.5%) to 97.3% (95% *CI* 97.0–97.6%) • Specificity 35.6% (95% *CI* 30.4–40.8%) to 99.5% (95% *CI* 99.4–99.6%) • No pooled analysis due to heterogeneity	103,018 (6, 14)	5.6 (2.9–9.8)	The estimated prevalence of the disease is between 1 and 4% and 10% over 70 years; the HF prevalence of the studies (case control excluded) is in this expected range The PPV ranged from 22.4 to 84.5%	For the patient selection domain, two studies had high risk of bias; one of them also applicability concerns. Another study had high risk of bias in the flow and timing domain. Poor reporting issues were found, especially concerning patient selection. One study did not report the number of diseased
Subgroup	Accuracy (95% CI)	No. of participants (studies, algorithms)	Prevalence Median (range)	Practical implication	Quality and comments
Hospitalized patients	Range • Sensitivity 29.0% (95% *CI* 27.4–30.6) to 96.0% (95% *CI* 91.0–99.0) • Specificity 84.3% (95% *CI* 84.0–84.6) to 99.2% (95% *CI* 98.8–99.6) • No pooled analysis due to heterogeneity	14,957 (8, 10)	13.7 (9.3–100)	There are not reliable estimates of the prevalence in unselected hospitalized population, which are however expected to be higher than in the outpatient setting, as is the case in the analyzed studies The PPV ranged from 35.8 to 94.0%	Six out of eight studies had quality concerns: for the patient selection domain, two studies had high and one study an unknown risk of bias, and an additional study had applicability concerns. Four studies had high risk of bias in the reference standard domain. Five studies had high risk of bias in the flow and timing domain

limiting our analysis to a single coding or health system or to a specific type of data such as hospital discharge records. Unfortunately, the low quality of the studies, especially concerning patient selection, together with applicability concerns for the same domain, did not allow to obtain a certain estimate of diagnostic accuracy of HAD-based algorithms for HF detection.

Conclusions

The considerable percentage of studies with a high risk of bias and the high clinical heterogeneity among different studies did not allow providing a pooled estimate of diagnostic accuracy for HAD-based algorithms intended for use in an unselected general adult population. Although the quality of the primary studies is low, excluding the study with the lowest sensitivity [46] and the one with the lowest specificity [38], algorithms applied in the general population have both sensitivities and specificities above 60%. However, to be able to obtain a correct summary estimate for the diagnostic accuracy of these algorithms, both in the unselected general population and in hospitalized subjects, a number of points should be addressed in future research. First, case–control designs (two-gates) should be avoided. It is fundamental to avoid verification bias caused by either preselecting subjects using a simpler version of the index test or verifying only subjects who are positive to the index [60]. Moreover, the new standard definition of HF [1] should be consistently applied, in order to have comparable spectra of diseased subjects across studies and countries. Finally, the reference standard should be applied in a more rigorous way, by either a prospective clinical evaluation or a standardized evaluation of all relevant clinical records by research-trained clinicians or nurses. More attention should be given to the purpose of the algorithm under development when determining participant inclusion criteria and choosing the validation setting. This translates into comprising a real unselected general population from census registry of an area if the aim of the algorithm is to determine HF prevalence. On the contrary, if the aim of the algorithm is to allow real-world studies on HF cohorts, strategies to increase specificity should be favored, such as the inclusion of high-prevalence populations and the use of multiple databases.

Abbreviations

ACCF/AHA	American College of Cardiology Foundation/American Heart Association
AF	Atrial fibrillation
AMI	Acute myocardial infarction
ATC	Anatomical therapeutic chemical
CA	Undergoing coronary angiography
CCS	Canadian Cardiovascular Society
CIHI	Canadian Institute for Health Information
COPD	Chronic obstructive pulmonary disease
DOR	Diagnostic odds ratio
EMR	Electronic medical records
ESC	European Society of Cardiology
ESS	Effective sample size
FN	False negative
FP	False positive
HAD	Health administrative data
HF	Heart failure
HT	Hypertension
ICD	International Classification of Diseases
MaRNet	Maritime Family Practice Research Network
NPV	Negative predictive value
OHIP	Ontario Health Insurance Plan
OPC	Outpatient clinic file
PPV	Positive predictive value
PTF	Patient treatment file
RCR	Review of clinical records
Sn	Sensitivity
Sp	Specificity
TN	True negative
TP	True positive
VA	Ventricular arrhythmia
VPOV	Veterans purpose of visit

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13643-024-02717-8.

Additional file 1. Data extraction form. Details of extracted variables and allowed values.

Additional file 2. Tailored QUADAS-2. Details of modified and added signalling questions.

Additional file 3. Full-text excluded with reason. List of excludes studies and detailed reason for exclusion.

Additional file 4. Additional information on included studies. Supplementary Table S1 – Reference standard: detailed definitions. Supplementary Table S2 – Details on included algorithms. Supplementary Table S3 – Diagnostic accuracy for studies including subgroups of general and hospitalized subpopulations. Supplementary Table S4 – Detailed QUADAS-2 evaluation. Supplementary Table S5 – PRISMA-DTA checklist. Supplementary Fig. S1 – Forest plot of positive predictive values (PPV) from studies in which only index test-positive subjects were verified with the reference standard. Supplementary Fig. S2 – Forest plots of sensitivity (Sn) and specificity (Sp) from studies including subgroups of the general population. Supplementary Fig. S3 – Forest plots of sensitivity (Sn) and specificity (Sp) from studies including subgroups of hospitalized patients with cardiovascular disease.

Acknowledgements

The authors would like to thank the University of Bicocca librarian, Laura Colombo, for the assistance with developing the literature search strategy. PROPHET-I (*PRediction Of chronic disease Progression through Health Electronic and Territorial data in Italian populations*) study group composition: Anita Andreano (1), Lucia Bisceglia (2), Martino Bussa (3), Caterina Fanizza (2), Vito Lepore (2), Pietro Magnoni (1), Cristina Mazzali (1), Alberto Milanese (1), Alessandro Musa (2), Giuseppe Occhino (2), Matteo Petrosino (2), Paola Rebora (3), Antonio Giampiero Russo (1), Deborah Testa (1), Maria Grazia Valsecchi (3), Adele Zanfino (1).
(1) Agenzia di Tutela della Salute della Città Metropolitana di Milano – SC Unità di Epidemiologia; (2) Agenzia Regionale Strategica per la Salute ed il Sociale (AReSS) Puglia – Area Epidemiologia e Care Intelligence; (3) Università Milano Bicocca – Dipartimento di Medicina e Chirurgia e Centro Interdipartimentale Bicocca Bioinformatics Biostatistics and Bioimaging Centre (B4).

Authors' contributions

PM, AM, and AA contributed to study design. Each author contributed to the protocol's planning and creation. PM and AM performed the bibliographic searches and screened titles and abstracts. AA and VL assessed full papers, extracted the data, performed data analysis, and drafted the manuscript. PR, LB, and AGR contributed to data interpretation. All authors revised and approved the final version of the manuscript. AGR acted as the guarantor of the review.

Funding

Funded by the European Union — Next Generation EU — NRRP M6C2 — Investment 2.1 Enhancement and strengthening of biomedical research in the NHS, project code PNRR-MAD-2022–12376033, title "Evidence-based models for high impact chronic disease prevention and risk of progression management in outpatient community services and community hospitals: towards eHealth integrating stratification on individual history with predictive models of disease progression, using machine learning and artificial intelligence on administrative and clinical databases," PI Antonio Giampiero Russo.

Data availability

The datasets underlying this article will be shared upon reasonable request to the corresponding author.

Declarations

Ethics approval and consent to participate
The study concerns literature-based studies. Therefore, ethical approval and informed consent were not required.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]SC Unità Di Epidemiologia, Agenzia Di Tutela Della Salute Della Città Metropolitana Di Milano, Via Conca del Naviglio 45, Milan 20123, Italy. [2]Area Epidemiologia E Care Intelligence, Agenzia Regionale Strategica Per La Salute Ed Il Sociale (AReSS) Puglia, Lungomare Nazario Sauro 33, Bari 70121, Italy. [3]Dipartimento Di Medicina E Chirurgia E Centro Interdipartimentale Bicocca Bioinformatics Biostatistics and Bioimaging Centre (B4), Università Milano Bicocca, Via Cadore 48, Monza 20900, Italy.

Received: 11 June 2024 Accepted: 18 November 2024
Published online: 24 December 2024

References
1. Bozkurt B, Coats AJS, Tsutsui H, et al. Universal definition and classification of heart failure: a report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure: endorsed by the Canadian Heart Failure Society, Heart Failure Association of India, Cardiac Society of Australia and New Zealand, and Chinese Heart Failure Association. Eur J Heart Fail. 2021;23(3):352–80. https://doi.org/10.1002/ejhf.2115.
2. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GMC, Coats AJS. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res. 2023;118(17):3272–87. https://doi.org/10.1093/cvr/cvac013.
3. Groenewegen A, Rutten FH, Mosterd A, Hoes AW. Epidemiology of heart failure. Eur J Heart Fail. 2020;22(8):1342–56. https://doi.org/10.1002/ejhf.1858.
4. Cowper DC, Hynes DM, Kubal JD, Murphy PA. Using administrative databases for outcomes research: select examples from VA Health Services Research and Development. J Med Syst. 1999;23(3):249–59. https://doi.org/10.1023/a:1020579806511.
5. Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. Annu Rev Public Health. 2011;32(1):91–108. https://doi.org/10.1146/annurev-publhealth-031210-100700.
6. Gini R, Francesconi P, Mazzaglia G, et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health. 2013;13:15. https://doi.org/10.1186/1471-2458-13-15.
7. Iron K, Lu H, Manuel D, Henry D, Gershon A. Using linked health administrative data to assess the clinical and healthcare system impact of chronic diseases in Ontario. Healthc Q Tor Ont. 2011;14(3):23–7. https://doi.org/10.12927/hcq.2011.22486.
8. Rector TS, Wickstrom SL, Shah M, et al. Specificity and sensitivity of claims-based algorithms for identifying members of Medicare+Choice health plans that have chronic medical conditions. Health Serv Res. 2004;39(6 Pt 1):1839–57. https://doi.org/10.1111/j.1475-6773.2004.00321.x.
9. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Lond Engl. 2020;396(10258):1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9.
10. Kopec JA. Estimating disease prevalence in administrative data. Clin Investig Med Med Clin Exp. 2022;45(2):E21-27. https://doi.org/10.25011/cim.v45i2.38100.
11. McCormick N, Lacaille D, Bhole V, Avina-Zubieta JA. Validity of heart failure diagnoses in administrative databases: a systematic review and meta-analysis. PLoS ONE. 2014;9(8): e104519. https://doi.org/10.1371/journal.pone.0104519.
12. Levinson RT, Malinowski JR, Bielinski SJ, et al. Identifying heart failure from electronic health records: a systematic evidence review. medRxiv. 2021;((Levinson R.T.) Clinic for General Internal Medicine and Psychosomatics, Heidelberg University Hospital, Heidelberg, Germany). https://doi.org/10.1101/2021.02.01.21250933.
13. Quach S, Blais C, Quan H. Administrative data have high variation in validity for recording heart failure. Can J Cardiol. 2010;26(8):306–12. https://doi.org/10.1016/s0828-282x(10)70438-4.
14. Lorenzoni G, Baldi I, Soattin M, Gregori D, Buja A. A systematic review of case-identification algorithms based on Italian Healthcare Administrative Databases for Three Relevant Diseases of the Cardiovascular System: hypertension, heart failure, and congenital heart diseases. Epidemiol Prev. 2019;43(4 Suppl 2):51–61. https://doi.org/10.19191/EP19.4.S2.P051.092.
15. Saczynski JS, Andrade SE, Harrold LR, Tjia J, Cutrona SL, Dodd KS, Goldberg RJ, Gurwitz JH. A systematic review of validated methods for identifying heart failure using administrative data. Pharmacoepidemiol Drug Saf. 2012;21(S1):129–40. https://doi.org/10.1002/pds.2313.
16. PROSPERO. https://www.crd.york.ac.uk/prospero/. Accessed 17 Apr 2024.
17. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282(11):1061–6. https://doi.org/10.1001/jama.282.11.1061.
18. Scherer RW, Saldanha IJ. How should systematic reviewers handle conference abstracts? A view from the trenches. Syst Rev. 2019;8(1):264. https://doi.org/10.1186/s13643-019-1188-0.
19. Szeto HC, Coleman RK, Gholami P, Hoffman BB, Goldstein MK. Accuracy of computerized outpatient diagnoses in a Veterans Affairs general medicine clinic. Am J Manag Care. 2002;8(1):37–43.
20. International Classification of Diseases (ICD). https://www.who.int/standards/classifications/classification-of-diseases. Accessed 9 Sept 2024.
21. Anatomical Therapeutic Chemical (ATC) Classification. https://www.who.int/tools/atc-ddd-toolkit/atc-classification. Accessed 9 Sept 2024.
22. Canadian Institute for Health Information | CIHI. https://www.cihi.ca/en. Accessed 7 May 2024.
23. Research, Statistics, Data & Systems | CMS. https://www.cms.gov/data-research. Accessed 7 May 2024.
24. Agency for Healthcare Research and Quality (AHRQ). https://www.ahrq.gov/. Accessed 7 May 2024.
25. Kohl C, McIntosh EJ, Unger S, et al. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. Environ Evid. 2018;7(1):8. https://doi.org/10.1186/s13750-018-0115-5.
26. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003;56(11):1129–35.
27. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36. https://doi.org/10.1059/0003-4819-155-8-201110180-00009.
28. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005;58(9):882–93. https://doi.org/10.1016/j.jclinepi.2005.01.016.
29. Macaskill, P, Gatsonis, S, Deeks, J, Harbord, R, Takwoingi,Y. Chapter 10: Analysing and presenting results. Cochrane Handb Syst Rev Diagn Test Accuracy Version 10 Cochrane Collab 2010. http://srdta.cochrane.org/. Accessed 31 July 2013.
30. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med. 1998;17(8):857–72.
31. Leeflang MM, Steingart KR, Scholten RJ, Davenport C. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (v2.0), Chapter 12 Drawing Conclusions; 2023.
32. Deeks, J, Bossuyt PM, Leeflang MM, Takwoingi,Y. Chapter 11.9 Presenting findings when meta-analysis cannot be performed. Cochrane Handb Syst Rev Diagn Test Accuracy Version 20. http://srdta.cochrane.org/. Accessed 31 July 2013.
33. Dunlay SM, Blecker S, Schulte PJ, Redfield MM, Ngufor CG, Glasgow A. Identifying patients with advanced heart failure using administrative

data. Mayo Clin Proc Innov Qual Outcomes. 2022;6(2):148–55. https://doi.org/10.1016/j.mayocpiqo.2022.02.001.

34. Vijh R, Wong ST, Grandy M, et al. Identifying heart failure in patients with chronic obstructive lung disease through the Canadian Primary Care Sentinel Surveillance Network in British Columbia: a case derivation study. CMAJ Open. 2021;9(2):E376–83. https://doi.org/10.9778/cmajo.20200183.
35. Xu Y, Lee S, Martin E, et al. Enhancing ICD-code-based case definition for heart failure using electronic medical record data. J Card Fail. 2020;26(7):610–7. https://doi.org/10.1016/j.cardfail.2020.04.003.
36. Cozzolino F, Montedori A, Abraha I, et al. A diagnostic accuracy study validating cardiovascular ICD-9-CM codes in healthcare administrative databases. The Umbria Data-Value Project. PloS One. 2019;14(7): e0218919. https://doi.org/10.1371/journal.pone.0218919.
37. Tison GH, Chamberlain AM, Pletcher MJ, et al. Identifying heart failure using EMR-based algorithms. Int J Med Inf. 2018;120:1–7. https://doi.org/10.1016/j.ijmedinf.2018.09.016.
38. Franchini M, Pieroni S, Passino C, Emdin M, Molinaro S. The CARPEDIEM algorithm: a rule-based system for identifying heart failure phenotype with a precision public health approach. Front Public Health. 2018;6:6. https://doi.org/10.3389/fpubh.2018.00006.
39. Schultz SE, Rothwell DM, Chen Z, Tu K. Identifying cases of congestive heart failure from administrative data: a validation study using primary care patient records. Chronic Dis Inj Can. 2013;33(3):160–6.
40. Alqaisi F, Williams LK, Peterson EL, Lanfear DE. Comparing methods for identifying patients with heart failure using electronic data sources. BMC Health Serv Res. 2009;9(1):237–237. https://doi.org/10.1186/1472-6963-9-237.
41. Teng THK, Finn J, Hung J, Geelhoed E, Hobbs M. A validation study: how effective is the Hospital Morbidity Data as a surveillance tool for heart failure in Western Australia? Aust N Z J Public Health. 2008;32(5):405–7. https://doi.org/10.1111/j.1753-6405.2008.00269.x.
42. So L, Evans D, Quan H. ICD-10 coding algorithms for defining comorbidities of acute myocardial infarction. BMC Health Serv Res. 2006;6:161. https://doi.org/10.1186/1472-6963-6-161.
43. Ingelsson E, Arnlöv J, Sundström J, Lind L. The validity of a diagnosis of heart failure in a hospital discharge register. Eur J Heart Fail. 2005;7(5):787–91. https://doi.org/10.1016/j.ejheart.2004.12.007.
44. Lee DS, Donovan L, Austin PC, Gong Y, Liu PP, Rouleau JL, Tu JV. Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. Med Care. 2005;43(2):182–8. https://doi.org/10.1097/00005650-200502000-00012.
45. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Med Care. 2005;43(5):480–5. https://doi.org/10.1097/01.mlr.0000160417.39497.a9.
46. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol. 2004;57(2):131–41. https://doi.org/10.1016/S0895-4356(03)00246-4.
47. Borzecki AM, Wong AT, Hickey EC, Ash AS, Berlowitz DR. Identifying hypertension-related comorbidities from administrative data: what's the optimal approach? Am J Med Qual Off J Am Coll Med Qual. 2004;19(5):201–6. https://doi.org/10.1177/106286060401900504.
48. Udris EM, Au DH, McDonell MB, et al. Comparing methods to identify general internal medicine clinic patients with chronic heart failure. Am Heart J. 2001;142(6):1003–9. https://doi.org/10.1067/mhj.2001.119130.
49. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. Ann Intern Med. 1993;119(8):844–50. https://doi.org/10.7326/0003-4819-119-8-199310150-00011.
50. Bosco-Lévy P, Duret S, Picard F, et al. Diagnostic accuracy of the International Classification of Diseases, Tenth Revision, codes of heart failure in an administrative database. Pharmacoepidemiol Drug Saf. 2019;28(2):194–200. https://doi.org/10.1002/pds.4690.
51. Kaspar M, Fette G, Güder G, et al. Underestimated prevalence of heart failure in hospital inpatients: a comparison of ICD codes and discharge letter information. Clin Res Cardiol. 2018;107(9):778–87. https://doi.org/10.1007/s00392-018-1245-z.
52. Blecker S, Katz SD, Horwitz LI, et al. Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data. JAMA Cardiol. 2016;1(9):1014–20. https://doi.org/10.1001/jamacardio.2016.3236.
53. Kümler T, Gislason GH, Kirk V, et al. Accuracy of a heart failure diagnosis in administrative registers. Eur J Heart Fail. 2008;10(7):658–60. https://doi.org/10.1016/j.ejheart.2008.05.006.
54. Quan H, Parsons GA, Ghali WA. Validity of information on comorbidity derived rom ICD-9-CCM administrative data. Med Care. 2002;40(8):675–85. https://doi.org/10.1097/00005650-200208000-00007.
55. Austin PC, Daly PA, Tu JV. A multicenter study of the coding accuracy of hospital discharge administrative data for patients admitted to cardiac care units in Ontario. Am Heart J. 2002;144(2):290–6. https://doi.org/10.1067/mhj.2002.123839.
56. P Macaskill, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Handbook for DTA Reviews | Diagnostic Test Accuracy Working Group. 2010. http://srdta.cochrane.org/handbook-dta-reviews. Accessed 30 Aug 2011.
57. Schmidt RL, Walker BS, Cohen MB. Verification and classification bias interactions in diagnostic test accuracy studies for fine-needle aspiration biopsy. Cancer Cytopathol. 2015;123(3):193–201. https://doi.org/10.1002/cncy.21503.
58. Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3):189–202. https://doi.org/10.7326/0003-4819-140-3-200402030-00010.
59. Saczynski JS, Andrade SE, Harrold LR, et al. A systematic review of validated methods for identifying heart failure using administrative data. Pharmacoepidemiol Drug Saf. 2012;21(S1):129–40. https://doi.org/10.1002/pds.2313.
60. Weiner MG, Garvin JH, Ten Have TR. Assessing the accuracy of diagnostic codes in administrative databases: the impact of the sampling frame on sensitivity and specificity. AMIA Annu Symp Proc AMIA Symp. 2006;2006:1140.

Publisher's Note