Lukas De Clercq*, Jelle C.L. Himmelreich and Ralf E. Harskamp

Quality of heart failure registration in primary care: observations from 1 million electronic health records in the Amsterdam Metropolitan Area

https://doi.org/10.1515/dx-2024-0009
Received January 10, 2024; accepted April 22, 2024;
published online May 14, 2024

Abstract

Objectives: Proper coding of heart failure (HF) in electronic health records (EHRs) is an important prerequisite for adequate care and research towards this vulnerable patient population. We set out to evaluate the accuracy of registration of HF diagnoses in primary care EHRs.
Methods: In a routine primary care database covering the Amsterdam Metropolitan Area, we identified all episodes of care with International Classification of Primary Care (ICPC) codes K77 (decompensatio cordis) or K84.03 (cardiomyopathy) up to 31/12/2021. We also performed two text-based searches to identify HF episodes without an appropriate ICPC-code. An expert panel evaluated all ICPC and text matches for congruence between the assigned codes and notes.
Results: From a database of 968,433 records we identified 19,106 patients (2.0 %) with a total of 24,011 ICPC-coded HF episodes. Removal of 1,324 episodes found to concern other or uncertain diagnoses and inclusion of 4,582 validated HF episodes identified through text search led to exclusion of 909 (overregistration: 4.8 %) and inclusion of 2,266 additional patients (underregistration: 11.1 %). The inclusion of miscoded HF episodes advanced the first known date of HF diagnosis in 3.9 % of records, with a median shift of 3.45 years. Episode-level underregistration decreased significantly over time, from 23.8 % in 2006 to 10.0 % in 2021.
Conclusions: While there is improvement over time, there are still substantial levels of over- and underregistration of HF, emphasizing the need for cautious interpretation of ICPC-coded data. The findings contribute to the understanding of HF registration issues in primary care and provide insights for improving registration practices.

Keywords: registration quality; primary care; heart failure; medical coding; medical records research

Introduction

Heart failure (HF) is a chronic disease that can lead to poor quality of life, high healthcare costs, and high mortality rates [1]. Moreover, with our aging populations we see a shift in the spectrum of HF and a growing proportion of patients being treated in primary care. Still most of our knowledge on HF, on symptoms, diagnosis, treatment and outcomes, stems from selected cohort studies and/or hospital care data [2]. It is likely that we hereby have an incomplete grasp of what HF entails, how it could be detected at an earlier stage, and managed, as well and its impact on our population.

Primary care electronic health records (EHRs) could provide valuable insights on this topic. These EHRs contain data on consultations and episodes as coded reason-for-encounter, diagnosis, or intervention(s). This process is standardized using the International Classification of Primary Care (ICPC). As healthcare settings differ between countries, various versions exist; in the Netherlands, general practitioners (GPs) use a tailored version called ICPC-1-NL [3, 4]. While these data may contain a wealth of data, concerns have been raised regarding the quality of registration. Previous studies have shown that quality of registration for clinical events coded using ICPC is suboptimal [5–8], with studies specifically on HF-coding in the Netherlands drawing the same conclusions [9, 10]. However, there are indications

***Corresponding author: Drs. Eng. Lukas De Clercq**, Department of General Practice, Amsterdam UMC location, University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands; and Personalized Medicine and Digital Health, Amsterdam Public Health, Amsterdam, The Netherlands, E-mail: l.declercq@amsterdamumc.nl. https://orcid.org/0000-0002-6175-9518
Jelle C.L. Himmelreich, Department of General Practice, Amsterdam UMC location, University of Amsterdam, Amsterdam, The Netherlands; Personalized Medicine, Amsterdam Public Health, Amsterdam, The Netherlands; and Heart Failure & Arrhythmias and Atherosclerosis & Ischemic Syndromes, Amsterdam Cardiovascular Sciences, Amsterdam, The Netherlands, E-mail: j.c.himmelreich@amsterdamumc.nl. https://orcid.org/0000-0003-0430-1583
Ralf E. Harskamp, Department of General Practice, Amsterdam UMC location, University of Amsterdam, Amsterdam, The Netherlands; Personalized Medicine, Amsterdam Public Health, Amsterdam, The Netherlands; and Heart Failure & Arrhythmias, Amsterdam Cardiovascular Sciences, Amsterdam, The Netherlands,
E-mail: r.e.harskamp@amsterdamumc.nl. https://orcid.org/0000-0001-9041-0350

that free text in patient records may be used to compensate for the shortfalls of taking ICPC-codes at face value for assessing presence of diagnoses [11].

In this study, we set out to study this hypothesis in which we validated episodes of care coded as HF in primary care patient records by evaluating the accompanying descriptions using an expert panel, allowing us to assess overregistration of HF. Conversely, for assessing underregistration, we identified un- or miscoded cases of HF using two text-based retrieval methods and validated the found matches.

Methods

Study design and setting

Our study used data from a large dynamic study cohort sourced from the Academic General Practice Research Network (AGPRN) of the Amsterdam University Medical Centers. Patients registered at over 100 practices in the Amsterdam Metropolitan Area affiliated with this network have the option to refuse reuse of their records for scientific research, though the opt-out rate is low (<0.5 %).

Participants

We included all patient records pertaining to patients who were at least 18 years of age and had one or more episodes of care registered on 31/12/2021.

HF identification

Patient records in our dataset were structured using episodes of care, which bundle consultation related to the same complaint or diagnosis. These episodes contain fields for ICPC-codes and physicians' notes, which we will capitalize on for this study. The database was searched for episodes labeled with K84.03 (cardiomyopathy), K77 (decompensation cordis), and the latter's subcodes K77.01 (acute decompensation cordis) and K77.02 (chronic decompensation cordis).

To investigate potential underregistration, we employed two information retrieval methods that make use of episode descriptions in order to find episodes that pertain to a HF diagnosis but are not coded as such.

The first search method makes use of manually engineered regular expressions (RegEx), a type of search pattern that facilitates complex query construction using wildcards, grouping, Boolean operators and many other features. Our designed expressions, which can be found in Supplement A, allow for some alternate or incorrect spellings of the search terms and exclude some expressions of uncertainty or exclusion of diagnosis. As the query is deterministic it yields a fixed number of matches.

The second text-based method further extends the idea of capturing incorrectly spelled terms or descriptions that pertain to HF by making use of FastText [12]. This algorithm converts words – preprocessed and separated into tokens – to vectorized representations. Similar to Word2Vec [13], these representations or "embeddings" attempt to capture semantic similarity, in which words with similar meaning have similar embeddings. In addition to this, FastText uses subword information to capture syntactic similarity, where words that share similar spelling will have similar embeddings. This subword embedding technique also allows for generating embeddings for out-of-vocabulary words. A FastText model was trained on all physician's notes in our dataset. Details regarding this training procedure can be found in Supplement B. Embeddings for episode description were generated by averaging across the vectorized representations of each word within [14]. These description embeddings were normalized for sequence length through dividing them by their L2-norm [15].

This would allow us to rank episodes by calculating the distance or similarity between their description embeddings and an embedding of a known reference representing a description for an HF episode. However, as descriptions of HF episodes may come in many forms, it doesn't suffice to look at the distance to a single reference. Our set of episodes with an HF ICPC-code contains a wide range of descriptions that we can employ as positive references, yet we lack a set of negatives. Covering the domain of descriptions that do not pertain to HF would require intractable amounts of manual validation and labeling, eliminating the typical binary classification algorithm options. The solution lies in fitting a one-class support vector machine (OCSVM) on the embedded descriptions of all validated HF-coded episodes [16]. Subsequently, all description embeddings of other episodes were ranked based on their distance from the decision boundary of the OCSVM. To facilitate comparison to the *k* descriptions found by the RegEx, we evaluate the top *k* episode descriptions as ranked by this method.

Validation

All episodes descriptions identified through either ICPC-code or textual search were evaluated by a panel of two reviewers. This panel represents expertise in general practice and cardiovascular morbidity, in both patient-facing as well as research roles (RH), combined with a background in medical informatics and the epidemiology of HF (LDC). Our study did not assess the validity of HF diagnoses themselves, but rather focused on the agreement between registered ICPC-codes and their accompanying descriptions. As such, only episodes that unambiguously excluded HF as a diagnosis, yet were coded as such, were considered to be false positives by our panel. This includes obvious accidental miscoding, explicit exclusion or uncertainty regarding an HF diagnosis, and diagnoses pertaining relatives of the patient. Inversely, in the evaluation of the textually retrieved episodes, our panel flagged only those that explicitly feature a HF diagnosis as positive cases. This approach provides the benefit of the doubt for both over- and underregistration and prioritizes specificity over sensitivity in the identification of miscoding. HF subtype information was registered when information regarding the left ventricular ejection fraction (LVEF) was found in the episode description, making a distinction between the HF subtypes with a reduced ejection fraction (HFrEF) for those with a LVEF<40 % and those with mildly-reduced (HFmrEF, LVEF=40–49 %) or preserved ejection fraction (HFpEF, LVEF≥50 %).

Statistical analysis

Over- and underregistration of HF: For our study, we defined overregistration rate as the false discovery rate (FDR) and underregistration rate as the false negative rate (FNR) of registered ICPC-codes for HF compared to their validated descriptions. Initially, we assessed these metrics at the episode level, where episode overregistration refers to the number of episodes incorrectly coded as HF (false positives) divided by the total number of episodes with such a code registered. Similarly, episode underregistration represents the number of episodes with a validated description of an HF diagnosis but lacking the appropriate ICPC-code (false negatives) divided by the total number of HF episodes, identified through ICPC-code or description. We plotted the annual over- and underregistration of HF episodes between 2006 and 2021 to assess their evolution over time. We tested for trend in the respective annual FDR- and FNR-ratios using the Mann-Kendall test [17, 18]. Confidence intervals (CIs) for these ratios were established using 10,000 bootstrapping iterations across all episodes in the denominators.

We then calculated these same metrics at the patient-level, where at least one validated HF episode in a patient record signified a positive and the lack thereof a negative HF case. This allowed us to assess the impact of over- and underregistration on the number of patient records retrieved when searching using ICPC-codes alone. We compared patient characteristics of the set of patients identified through ICPC-alone to those of a set in which the detected over- and underregistration was corrected. These characteristics were: age at time of first HF diagnosis, patient sex, and a set of HF risk factors derived from literature in previous work, identified through ICPC-coded episodes [19]. Age, being a continuous variable, was tested for independence using a Mann-Whitney U test [20]. The remaining categorical values were subjected to a chi-squared test for independence [21]. The significance level was held at 0.05, with a Bonferroni correction applied for multiple comparison. In addition, we investigated delayed coding by identifying the number of records in which the first-known HF diagnosis date was adjusted based on the newly identified, miscoded HF episodes. We measured the extent of the adjustment for these patient records and express it as median and first and third quartiles.

Distribution of ICPC-codes in miscoded HF episodes: To investigate potential causes of underregistration, we analyzed the distribution of ICPC-codes that were registered for the false negative HF episodes, i.e. those without an appropriate HF code. To identify opportunities for compensating suboptimal coding, we analyzed the two subsets of false negative episodes where the miscoding had an impact at the patient-level. The first subset was that of patients with no other validated HF episodes, while the second subset included episodes that led to an earlier first known HF date.

HF subtype registration prevalence: Our findings regarding the identification of HF subtype – as distinguished by LVEF – were described at the episode-level, aiming to provide a look into the level of detail of HF episode descriptions by assessing the prevalence at which this subclassification was provided by the attending GP.

Comparison of text retrieval methods: We compared the results of our two free-text information retrieval methods, RegEx and OCSVM, using the precision metric, representing the fraction of unique retrieved HF descriptions that were validated as such. As the results for the OCSVM are ranked, this metric is represented by precision-at-*k*, measuring performance across the top *k* predictions. These metrics are calculated for each method individually, their union, and their intersection (overlap).

Results

Over- and underregistration of HF

A total of 19,256,707 episodes were queried for HF using both ICPC-codes and descriptions, spanning 968,433 EHRs. The results of validating these codes against their descriptions can be seen in Table 1.

At the episode-level, 24,011 carried an HF ICPC-code, the majority of which were found under the K77 code (87.4 %). Our panel identified 1,324 as false positive, an overregistration rate of 5.5 %. Using the validated episode descriptions found in our textual search, we identified an additional 4,582 HF episodes, which translates to an underregistration rate of 16.8 %. Plotting these measurements over time, we observed a significant decrease in the HF episode underregistration rate (p<0.001), going from 23.8 % (95 % CI: 21.5–26.5) in 2006 to 10.0 % (95 % CI: 8.3–11.1) in 2021, as can be seen in Figure 1. By way of contrast, the overregistration rate did not present a significant trend (p=0.096).

In terms of patient records, these values translate to 19,106 patients with an HF ICPC-coded episode, of which 909 records are eliminated once episodes incorrectly coded as HF were removed, making for an overregistration rate of 4.8 %. Conversely, an additional 2,266 records of patients with a HF diagnosis were identified using the additional HF episodes found by the textual search, translating to a patient-level underregistration rate of 11.1 %. The result of this effort

Table 1: Contingency tables of registered heart failure validation.

(A) Episodes		Coded			(B) Patients		Coded		
		Positive	Negative				Positive	Negative	
Validated	Positive	22,687	4,582	FNR=16.8 %	Validated	Positive	18,197	2,266	FNR=11.1 %
	Negative	1,324	19,228,114			Negative	909	947,061	
		FDR=5.5 %					FDR=4.8 %		

Contingency tables comparing ICPC-coding for heart failure against a validated reference at the episode-level (A) and the patient-level (B). Overregistration is expressed as false discovery rate (FDR), underregistration as false negative rate (FNR). Values in the lower right quadrant are depicted in gray to indicate that they contain unvalidated episodes.

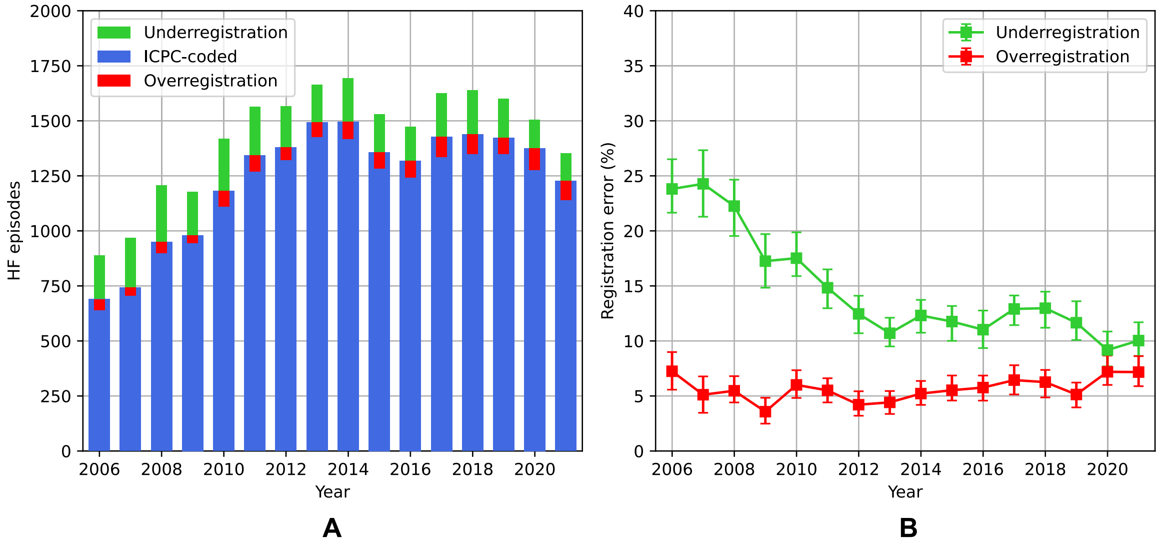

Figure 1: Heart failure registration quality over time. Under- and overregistration of heart failure (HF) episodes per year (A) and expressed as rate in reference to the number of validated HF episodes (B). Stems indicate the 95 % confidence interval of registration error rates.

was a total of 20,463 EHRs validated as having HF (2.1 %). After inclusion of the previously unidentified HF episodes, the first known date of HF diagnosis was determined to be earlier in 717 records (3.7 %), with a median shift of 3.45 years (IQR 0.97–8.11). The effects of these false and missed inclusions on population characteristics are apparent in Table 2. We observed a statistically significant shift in the patients' age and history of hypertension, coronary artery disease, atrial fibrillation, and chronic kidney disease, as registered at the time of HF diagnosis.

Table 2: Heart failure population characteristics shift due to registration errors.

	ICPC-coded (n=19,106)		Post-correction (n=20,463)		p-Value
	Median	**25th – 75th**	**Median**	**25th – 75th**	
Age at HF diagnosis, years	75.88	(65.00–83.93)	75.31	(64.12–83.63)	<0.001[a]
	n	**(%)**	**n**	**(%)**	
Male sex	9,065	(47.4)	9,818	(48.0)	0.293
Hypertension	8,074	(42.3)	8,151	(39.8)	<0.001[a]
Coronary artery disease	5,378	(28.1)	5,358	(26.2)	<0.001[a]
Diabetes mellitus	4,947	(25.9)	5,084	(24.8)	0.017
Atrial fibrillation	4,012	(21.0)	3,897	(19.0)	<0.001[a]
Chronic obstructive pulmonary disease	2,903	(15.2)	2,915	(14.2)	0.008
Chronic kidney disease	2,645	(13.8)	2,575	(12.6)	<0.001[a]
Valvular heart disease	2,044	(10.7)	2,014	(9.8)	0.005
Stroke	1,805	(9.4)	1,810	(8.8)	0.039
Obesity	1,374	(7.2)	1,353	(6.6)	0.024
Tobacco use	1,160	(6.1)	1,168	(5.7)	0.130
Alcohol abuse	691	(3.6)	712	(3.5)	0.478
Heart murmur	159	(0.8)	164	(0.8)	0.777
Material deprivation	110	(0.6)	111	(0.5)	0.706
Family history of cardiovascular disease	81	(0.4)	82	(0.4)	0.778

Comparison of population demographics and heart failure (HF) risk factor prevalence between the set identified through ICPC-code search and that after in- and exclusion of identified miscoded episodes. All characteristics were determined at the time of first known HF diagnosis. [a]p-Values were found to be significant at the 0.05 level after Bonferroni correction.

Distribution of ICPC-codes in miscoded HF episodes

Table 3 presents the ICPC-codes most frequently found across all miscoded HF episodes. These rates are also expressed for the subsets of “impactful” episodes, i.e. those that yielded either new patient records with HF or those that precede the first-known HF ICPC-code in a record, as opposed to those that follow a correctly ICPC-coded HF episode. Out of all episodes which ought to have a HF ICPC-code but did not, more than a quarter (25.3 %) lacked any ICPC-code. More than half (55.8 %) were found in ICPC’s chapter K, encompassing cardiovascular afflictions, which reached 59.3 % of coverage when only considering impactful episodes. The most prevalent code among these was K78 (Atrial fibrillation), making up 11.7 % of all newly identified episodes and 13.2 % of impactful episodes. The top ten ICPC-codes were responsible for nearly half (48.7 %) of impactful episodes and all but one, R02 (Dyspnea attributed to respiratory system), fall in chapter K. The distribution across the different chapters may be found in Supplemental Table C.

HF subtype registration

Our panel was able to identify 2,532 episodes (11.2 %) describing HFrEF and 1,558 episodes (6.9 %) pertaining to HFmrEF or HFpEF. The remaining 18,597 episodes (82.2 %) did not name the subtype and contained no information regarding LVEF.

Identification of HF diagnosis descriptions

The RegEx filters identified a total of k=4,881 unique episode descriptions, of which 3,260 were verified as pertaining to HF by the panel, yielding a precision of 66.8 % for this manually defined method. Evaluation of the top k as ranked by the OCSVM yielded 2,391 HF episode descriptions validated as such, for a precision-at-k of 49.0 % for this self-supervised method. A visual comparison of these precisions can be found in Supplemental Figure D. The union of both methods yielded a total of 7,518 unique descriptions, of which 3,761 were validated as pertaining to HF, for a total free-text search precision of 50.0 %. A subset of 2,244 descriptions was identified by both retrieval methods, of which 1,890 were validated as HF, achieving a precision of 84.2 % for the intersection of both methods.

Discussion

This study highlighted the challenges in accurately identifying and coding HF episodes in primary care EHRs. Our

Table 3: Distribution of miscoded heart failure episodes per ICPC-code.

ICPC-code		All (%)		Novel (%)		Preceding (%)		Impactful (%)	
–	No registered code	1,157	(25.3)	651	(28.7)	135	(18.8)	786	(27.8)
K78	Atrial fibrillation	534	(11.7)	290	(12.8)	105	(14.6)	395	(14.0)
K84	Heart disease (other)	375	(8.2)	195	(8.6)	47	(6.6)	242	(8.6)
K75	Acute myocardial infarction	353	(7.7)	198	(8.7)	81	(11.3)	279	(9.9)
K87	Hypertension complicated	162	(3.5)	72	(3.2)	34	(4.7)	106	(3.7)
K76	Ischemic heart disease (without angina)	150	(3.3)	62	(2.7)	25	(3.5)	87	(3.1)
R02	Dyspnea attributed to respiratory system	113	(2.5)	57	(2.5)	21	(2.9)	78	(2.8)
K02	Pressure/tightness attributed to heart	98	(2.1)	57	(2.5)	14	(2.0)	71	(2.5)
K74	Angina pectoris	96	(2.1)	51	(2.3)	18	(2.5)	69	(2.4)
K83	Heart valve disease (non-rheumatic)	89	(1.9)	47	(2.1)	19	(2.6)	66	(2.3)
K86	Hypertension uncomplicated	88	(1.9)	48	(2.1)	12	(1.7)	60	(2.1)
U99.01	Renal insufficiency	77	(1.7)	9	(0.4)	13	(1.8)	22	(0.8)
K07	Peripheral edema	72	(1.6)	32	(1.4)	10	(1.4)	42	(1.5)
K83.02	Mitral regurgitation	72	(1.6)	23	(1.0)	18	(2.5)	41	(1.5)
R81	Pneumonia	65	(1.4)	27	(1.2)	7	(1.0)	34	(1.2)
K82	Pulmonary heart disease	61	(1.3)	37	(1.6)	8	(1.1)	45	(1.6)
K83.01	Aortic stenosis	61	(1.3)	33	(1.5)	13	(1.8)	46	(1.6)
K99	Cardiovascular disease other	50	(1.1)	18	(0.8)	6	(0.8)	24	(0.8)
A05	General deterioration/feeling ill	47	(1.0)	10	(0.4)	11	(1.5)	21	(0.7)
A89.01	Pacemaker/internal defibrillator	43	(0.9)	12	(0.5)	4	(0.6)	16	(0.6)
	Other	819	(17.9)	337	(14.9)	116	(16.2)	453	(15.2)

Distribution of miscoded heart failure (HF) episodes across the most frequent ICPC-codes. “Novel” indicates HF episodes in records with no validated coded HF, “Preceding” refers to those that advanced the earliest known diagnosis of HF in a patient record, “Impactful” is the combination of both.

assessment revealed substantial levels of HF over- and underregistration in routine primary care EHRs in the Amsterdam Metropolitan Area, with underregistration affecting a larger share of patients. This shows that taking registered HF ICPC-codes at face value would incorrectly include and exclude large fractions of patients. Furthermore, we have shown that suboptimal labeling of HF episodes can also impact the first-known date of HF diagnosis, which in turn may introduce bias in various descriptive and predictive analyses. The observed shifts in demographics and risk factors indicate that epidemiological HF research or modeling efforts may be adversely impacted if HF miscoding is unaccounted for. In daily practice, miscoding provides friction in the evaluation of a patient record as well as hinder automated alerts and decision support embedded in the HER, and may continue to do so as these records are transferred between practices.

We observed a significant downward trend for underregistration of HF. One possible driver of this improvement is provided by the Netherlands Institute for Health Services Research (NIVEL) in the form of regular feedback provided to GPs on the quality of their registration [22, 23]. Other initiatives that might have contributed to this trend include updated registration quality guidelines emphasizing episode-oriented registration [24], financial incentives tied to quality measures [23], and continuing education such as e-learning courses on HF and record-keeping provided by the Dutch College of General Practitioners (NHG) [25, 26]. By means of contrast, there was insufficient evidence to identify such a trend for overregistration. In the evaluation of episodes with an HF ICPC-code our panel identified a non-trivial amount of episode descriptions as false positives as they expressed uncertainty regarding the diagnosis of HF or exclusion thereof. The ICPC does not allow for flagging of degrees of certainty, although proposals along these lines have been made in the past [27]. Overregistration of HF may remain an issue due to GPs registering preliminary diagnoses or the exclusion thereof under a HF ICPC-code rather than the most applicable synonym. Further exploration of this phenomenon is required to identify mitigation strategies.

We recommend that future HF research efforts on similar datasets take the observed deficiencies into account. To this end, have identified several strategies to compensate for miscoding. Episodes pertaining to HF found using their description were heavily concentrated in a handful of ICPC-codes, many of which were strongly related to HF. The prevalence at which HF diagnoses were found under the code for atrial fibrillation, a condition frequently diagnosed simultaneous with HF, may indicate that some GPs add both in a single episode. Finally, combining episodes lacking an ICPC-code and those under ICPC's chapter K covered over 85 % of the additionally found HF patients. This may provide a substantially reduced search space in future work that aims to include miscoded HF.

Related work

Validation of ICPC-coding in the Netherlands

Several studies have investigated the accuracy of ICPC-coding in the Netherlands, with a wide range of results across the different conditions they investigated.

In a validation study of a sample (n=200) from IPCI, a Dutch primary care database similar to that of the AGPRN, Coloma et al. identified a “best-case” overregistration at the patient-level of 25.0 % for those with an ICPC1-NL code for acute myocardial infarction (AMI). When they included records deemed non-assessable in the denominator (the “worst-case” scenario), the overregistration reached 63.5 %. Based on these results, they recommend complementing the use of ICPC-codes with free-text search [7].

The use of free-text search to compensate for ICPC quality issues was also touched upon by Valkhoff et al. in their study of the quality of gastrointestinal bleeding overregistration. In their sample (n=200) the patient-level FDR of relevant ICPC-codes (78 %) was equal to that of a free-text search (79 %). They concluded that textual retrieval is valuable for identifying additional cases, yet should not be used without manual validation [6].

In a select sample (n=6,671) of patients with a body mass index of ≥27 kg/m^2, de Boer et al. found the overregistration rate of coded diabetes mellitus (DM) at the patient-level to be negligible (0.02 %). Underregistration of DM was higher, with a patient-level FNR of 9.1 %. It is to be noted that the GPs were aware of this study prior to extraction, which may have influenced the registration quality of the patient records [8].

By linking the Netherlands Cancer Registry with primary care EHRs of 290,000 patients, Sollie et al. were able to estimate the registration quality of cancer. Nearly half of all episodes coded for cancer were not found in the registry, translating to a patient-level overregistration rate of 48.9 %. A little over half of those registered in the NCR had the corresponding code in their record, resulting in an underregistration rate of 39.4 % [5].

Over-/underregistration of HF in primary care

A database study in Switzerland found a coded HF prevalence of 0.5 % in GP records between 2016 and 2019 (n=1,288). Their findings contrasted with the national prevalence estimate of 2.5 %, indicating a serious underdiagnosing or -registration of HF. Rachamin et al. suggest both the high coverage of ambulatory cardiologists and inconsistent use of coding by GPs as causes for this discrepancy [1].

Raat et al. investigated over- and underdiagnosis and -registration of HF in Belgium using a select sample (n=4,678) of patients over 40 years old with one or more registered risk factors or prescriptions related to HF. All patients were re-assessed by their GPs for presence of HF. They found a patient-level underregistration rate of 69.5 % and overregistration rate of 47.1 % [9].

In the Netherlands, HF overregistration was studied by Valk et al., who conducted a cross-sectional study of 683 patients in 30 general practices in the Amersfoort region with a registered ICPC-code for HF (K77). Their panel was able to confirm the HF diagnosis for 434 of the patient records. In 118 of cases the panel considered HF to be absent and for 131 the diagnosis remained uncertain, translating to an FDR at the patient-level of 17.3–36.5 %. This study did not assess underregistration [10].

HF subtype registration

The 2021 publication and subsequent wide endorsement of the "Universal Definition and Classification of Heart Failure" as a collaboration between the Heart Failure Association of the European Society of Cardiology (HFA-ESC), the Heart Failure Society of America (HFSA), and the Japanese Heart Failure Society (JHFS) is indicative of the growing need for improved registration of HF. This guideline has an emphasis on the differentiation between LVEF subgroups due to its prognostic qualities, its ability to identify groups known to respond to life-prolonging treatment, and its familiarity outside the context of the cardiology specialty [28].

In response to this growing need, the NHG published a new version of the ICPC-1-NL in 2022 [29]. This update includes two new subcodes intended to replace *chronic decompensation cordis* (K77.02): HFpEF (K77.03) and HFrEF/HFmrEF (K77.04). This paves the way for improved registration, provided these codes are integrated into EHRs and GPs are adequately trained in applying them. A similar proposal was accepted for ICPC-3 [30].

Strengths and limitations

The scale and coverage of our dataset instils confidence regarding its representativeness of the population in the Amsterdam Metropolitan Area. Whether our findings translate to the Netherlands as a whole or other countries employing ICPC-coding remains to be seen.

In comparison to most validation studies of ICPC-coding in the Netherlands, our sample size is substantially bigger. This allowed us to investigate HF underregistration, being the first to do so in the Netherlands, as well as the distribution of ICPC-codes in the additionally found HF episodes.

Our information retrieval method is partially based on FastText, which generates static word embeddings that do not take into account the contexts in which words are found in. As such, the OCSVM's performance was hampered by its inability to take e.g. statements of negation or doubt into account. Even though this was partially alleviated by combining its results with those of the RegEx-based search, this did not reduce the number of descriptions to be validated and it is likely that our retrieval methods missed miscoded HF episodes where their shortcomings overlap. More recent language models produce contextualized embeddings and are likely to substantially improve both precision and recall of this retrieval task. These can come in the form of general purpose language models such as LLaMA [31] or GPT-4 [32], or Dutch medical domain-specific models such as MedRoBERTa.NL [33].

In the validation of retrieved episodes, data privacy regulations prevented the use of an independent panel, introducing a risk for validation bias. To compensate, our panel remained conservative in its judgments. Combined with a lack of sufficient detail in many descriptions, this likely prevented identification of certain false positives and false negatives. As such, the numerators of both the over- and underregistration rates are likely to be underestimates and ought to be interpreted as lower limits. Furthermore, the correlation between the elaborateness of an episode description and the level of scrutiny that allows for may have introduced a bias in favor of less adequate notes.

Validation efforts being hampered by a large variety in note-taking quality and quantity is not a new phenomenon. Similar conclusions were drawn by Sporaland et al. in their investigation into the congruence between ICPC2-codes and their accompanying descriptions [34]. Future efforts may seek to assess an entire patient record in the validation of a code, looking beyond the episode it's attached to, as has been performed by panels on smaller samples in the past [6, 7, 10]. Information from subsequent consultations, referrals, and prescriptions is likely to substantially improve the detection of miscoding. However, such a validation task may be considered intractable when considering datasets of this scale. Solace may be found in the aforementioned advancements in language models, whose contextual and generative capacities may be leveraged for such an effort. In this pursuit, a smaller sample may be evaluated concurrently by a panel to assess the validation performance of such a language model before applying it to broader datasets.

Finally, the use of the designated episode start date rather than the date at which it was registered is a trade-off, as GPs may choose to antedate HF diagnoses where they deem applicable, yet registration dates were found to contain technical inaccuracies in this particular dataset. Our choice in this trade-off may have introduced bias in the temporal analysis.

Conclusions

Our findings revealed significant levels of both over- and underregistration of HF and its subtypes in primary care EHRs. The observed biases resulting from such miscoding indicate that relying solely on registered ICPC-codes can lead to erroneous epidemiological research and modeling. Analysis of the trends over time and the distribution of these errors point to potential shortcomings in the ICPC-1-NL system and coding practices employed by GPs in the Netherlands. Free-text search and information retrieval methods combined with a targeting of the subset of ICPC-codes these HF coding errors are concentrated in may provide an avenue to mitigate these shortcomings in EHR database research.

Research ethics: The study was authorized by the local Institutional Review Board (Amsterdam UMC Medical Ethics Committee, W23_07#23.097).
Informed consent: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: This work was supported by departmental resources.
Data availability: Requests for access to the raw data are subject to terms of service and privacy regulations and can be directed at the Amsterdam UMC General Practice Network (datamanagementanha@amsterdamumc.nl) using the reference “Target-HF”.

References

1. Rachamin Y, Meier R, Rosemann T, Flammer AJ, Chmiel C. Heart failure epidemiology and treatment in primary care: a retrospective cross-sectional study. ESC Heart Fail 2021;8:489–97.
2. Hobbs FDR, Doust J, Mant J, Cowie MR. Diagnosis of heart failure in primary care. Heart 2010;96:1773–7.
3. Lamberts H, Wood M. ICPC, international classification of primary care. USA: Oxford University Press; 1987.
4. HIS-Referentiemodel & NHG-Tabellen. Nederlands Huisartsen Genootschap. Available from: https://referentiemodel.nhg.org/ [Accessed 7 Jun 2023].
5. Sollie A, Roskam J, Sijmons RH, Numans ME, Helsper CW. Do GPs know their patients with cancer? Assessing the quality of cancer registration in Dutch primary care: a cross-sectional validation study. BMJ Open 2016;6:e012669.
6. Valkhoff VE, Coloma PM, Masclee GM, Gini R, Innocenti F, Lapi F, et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. J Clin Epidemiol 2014;67: 921–31.
7. Coloma PM, Valkhoff VE, Mazzaglia G, Nielsson MS, Pedersen L, Molokhia M, et al. Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries. BMJ Open 2013;3: e002862.
8. de Boer A, Blom J, de Waal M, Rippe R, de Koning E, Jazet I, et al. Coded diagnoses from general practice electronic health records are a feasible and valid alternative to self-report to define diabetes cases in research. Prim Care Diabetes 2021;15:234–9.
9. Raat W, Smeets M, Henrard S, Aertgeerts B, Penders J, Droogne W, et al. Machine learning optimization of an electronic health record audit for heart failure in primary care. Esc Heart Fail 2022;9:39–47.
10. Valk MJ, Mosterd A, Broekhuizen BD, Zuithoff NP, Landman MA, Hoes AW, et al. Overdiagnosis of heart failure in primary care: a cross-sectional study. Br J Gen Pract 2016;66:e587–92.
11. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inf Assoc 2014;21:221–30.
12. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Ling 2017;5:135–46.
13. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.
14. Iyyer M, Manjunatha V, Boyd-Graber J, Daumé IIIH. Deep unordered composition rivals syntactic methods for text classification. Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (vol 1: Long papers); Bejing, China; Association for Computational Linguistics; 2015.
15. Wang Y, Chao W-L, Weinberger KQ, van der Maaten L. Simpleshot: revisiting nearest-neighbor classification for few-shot learning. arXiv preprint arXiv:191104623. 2019.
16. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural Comput 2001;13:1443–71.
17. Mann HB. Nonparametric tests against trend. Econometrica 1945;13(3): 245–59.
18. Kendall MG. Rank correlation methods. Griffin; 1948.
19. De Clercq L, Schut MC, Bossuyt PM, van Weert HC, Handoko ML, Harskamp RE. TARGET-HF: developing a model for detecting incident heart failure among symptomatic patients in general practice using routine health care data. Fam Pract 2023;40:188–94.
20. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947; 18(1):50–60.

21. Ludbrook J. Analysis of 2 × 2 tables of frequencies: matching test to experimental design. Int J Epidemiol 2008;37:1430–5.
22. Jabaaij L, Verheij R, Njoo K, Hoogen HVD, Tiersma W, Levelink H. Het meten van de kwaliteit van de registratie in elektronische patiënten dossiers van huisartsen met de EPD-scan-h. Utrecht: NIVEL; 2008.
23. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. J Med Internet Res 2018;20:e185.
24. Duineveld B, Kole HM, Van Werven H, Sloekers J, Njoo KH. Richtlijn Adequate dossiervorming met het EPD (ADEPD) versie 4: Nederlands Huisartsen Genootschap; 2019. Available from: https://www.nhg.org/praktijkvoering/informatisering/richtlijn-adequate-dossiervorming-epd/ [Accessed 8 Jun 2023].
25. NHG E-learning Hartfalen: Nederlands Huisartsen Genootschap. Available from: https://www.nhg.org/product/hartfalen/ [Accessed 8 Jun 2023].
26. NHG E-learning ADEPD: Nederlands Huisartsen Genootschap. Available from: https://www.nhg.org/product/adepd/ [Accessed 8 Jun 2023].
27. Napel HT, Boven KV. ICPC-3 International Classification of Primary Care: user manual and classification. ISBN 9781032053394; CRC Press, Boca Raton; 2022
28. Bozkurt B, Coats AJ, Tsutsui H, Abdelhamid M, Adamopoulos S, Albert N, 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. J Card Fail 2021;27:387–413.
29. Advies voor doorvoeren wijzigingen subcodes bij NHG-Tabel 24 - ICPC, versie 10: Nederlands Huisartsen Genootschap; 2022. Available from: https://referentiemodel.nhg.org/sites/default/files/NHG-Tabel%2024-ICPC-versie-10-Advies%20doorvoeren%20wijzigingen.pdf [Accessed 12 Jun 2023].
30. ICPC-3 update meeting: WONCA-ICPC Foundation; 2022. Available from: https://icpc-3.info/documents/extra/proposals2022.pdf [Accessed 12 Jun 2023].
31. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. LLaMa: open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.
32. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.
33. Verkijk S, Vossen P. MedRoBERTa.nl: a language model for Dutch electronic health records. Comput Ling Neth J 2021;11:141–59.
34. Sporaland GL, Mouland G, Bratland B, Rygh E, Reiso H. General practitioners' use of ICPC diagnoses and their correspondence with patient record notes. Tidsskr Nor Legeforen 2019;139(15). https://doi.org/10.4045/tidsskr.18.0440.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/dx-2024-0009).