


Machine learning optimization of an electronic health record audit for heart failure in primary care

Willem Raat^{1*} , Miek Smeets¹, Severine Henrard², Bert Aertgeerts¹, Joris Penders^{3,4}, Walter Droogne⁵, Wilfried Mullens^{3,4}, Stefan Janssens⁵ and Bert Vaes¹

¹Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium; ²Louvain Drug Research Institute, Clinical Pharmacy Research Group (CLIP) and Institute of Health and Society (IRSS), Université catholique de Louvain (UCLouvain), Brussels, Belgium; ³Ziekenhuis Oost-Limburg, Genk, Belgium; ⁴University of Hasselt, Hasselt, Belgium; and ⁵Department of Cardiovascular Diseases, University Hospitals Leuven, KU Leuven, Leuven, Belgium

Abstract

Aims The diagnosis of heart failure (HF) is an important problem in primary care. We previously demonstrated a 74% increase in registered HF diagnoses in primary care electronic health records (EHRs) following an extended audit procedure. What remains unclear is the accuracy of registered HF pre-audit and which EHR variables are most important in the extended audit strategy. This study aims to describe the diagnostic HF classification sequence at different stages, assess general practitioner (GP) HF misclassification, and test the predictive performance of an optimized audit.

Methods and results This is a secondary analysis of the OSCAR-HF study, a prospective observational trial including 51 participating GPs. OSCAR used an extended audit based on typical HF risk factors, signs, symptoms, and medications in GPs' EHR. This resulted in a list of possible HF patients, which participating GPs had to classify as HF or non-HF. We compared registered HF diagnoses before and after GPs' assessment. For our analysis of audit performance, we used GPs' assessment of HF as primary outcome and audit queries as dichotomous predictor variables for a gradient boosted machine (GBM) decision tree algorithm and logistic regression model. Of the 18 011 patients eligible for the audit intervention, 4678 (26.0%) were identified as possible HF patients and submitted for GPs' assessment in the audit stage. There were 310 patients with registered HF before GP assessment, of whom 146 (47.1%) were judged not to have HF by their GP (over-registration). There were 538 patients with registered HF after GP assessment, of whom 374 (69.5%) did not have registered HF before GP assessment (under-registration). The GBM and logistic regression model had a comparable predictive performance (area under the curve of 0.70 [95% confidence interval 0.65–0.77] and 0.69 [95% confidence interval 0.64–0.75], respectively). This was not significantly impacted by reducing the set of predictor variables to the 10 most important variables identified in the GBM model (free-text and coded cardiomyopathy, ischaemic heart disease and atrial fibrillation, digoxin, mineralocorticoid receptor antagonists, and combinations of renin-angiotensin system inhibitors and beta-blockers with diuretics). This optimized query set was enough to identify 86% ($n = 461/538$) of GPs' self-assessed HF population with a 33% reduction ($n = 1537/4678$) in screening caseload.

Conclusions Diagnostic coding of HF in primary care health records is inaccurate with a high degree of under-registration and over-registration. An optimized query set enabled identification of more than 80% of GPs' self-assessed HF population.

Keywords Chronic heart failure; Primary care; Audit and feedback; Electronic health records; Screening

Received: 19 July 2021; Revised: 27 October 2021; Accepted: 6 November 2021

*Correspondence to: Willem Raat, Department of Public Health and Primary Care, KU Leuven, Kapucijnenvoer 33, blok j bus 7001, 3000 Leuven, Belgium. Tel: +32 (0)16 337 468. Email: willem.raat@kuleuven.be

Introduction

Heart failure (HF) is an important and growing health concern.¹ Patients are mainly older and fragile. Thus, general practitioners (GPs) are ideally placed to deliver care.²

However, the identification of HF in this older primary care population is difficult because signs and symptoms are non-specific.³ This can lead to overdiagnosis and underdiagnosis^{4,5} and suboptimal diagnostic coding in the electronic health record (EHR),³ which is associated with

poorer outcomes in HF and other chronic illnesses.^{6–8} Because data on diagnostic coding are easily extractable from the EHR, it is a particularly suitable target for audit and feedback strategies in primary care. These are based on the belief that healthcare professionals are prompted to modify their practice when given performance feedback showing that their clinical practice is inconsistent with a desirable target. They have been associated with small but potentially important improvements in professional practice.⁹

We previously designed and reported a multifaceted intervention in eight general practices ($n = 18\ 011$ patients ≥ 40 years) aimed at improving case-finding through an audit in GPs' EHR.^{10,11} Although the audit resulted in a more than 74% increase in the number of registered HF diagnoses in GPs' EHR, the accuracy of registered HF pre-audit remains unclear. In addition, the audit generated a list of possible HF patients of whom a large majority did not have HF, which significantly increased GPs' time investment and could therefore limit implementation in a real-world setting.

The aim of this paper is two-fold: first, to describe the diagnostic HF classification sequence at three stages, namely, pre-audit, GP assessment, and panel validation, and assess the magnitude of GP HF misclassification; and, second, to analyse the classification performance of the query-based audit in the EHR and test the predictive performance of an optimized audit.

Methods

Design and setting

We previously reported on the design and methods used in the OSCAR-HF study.¹¹ This was a non-randomized, non-controlled prospective observational trial with 6 months of follow-up conducted in eight Belgian general practices in

2017, including 51 participating GPs, and piloting the use of a basic and extended audit in the EHR. The basic audit queried registered HF diagnoses (coded or free text). The extended audit queried several coded and free-text search strings mapping known HF risk factors such as signs, symptoms, comorbidities, as well as typical HF medications (*Table 1*). These two audits resulted in a list of possible HF patients, which participating GPs then classified as HF or non-HF. Patients classified as having HF constituted the OSCAR-HF study population.¹⁰ In addition, an expert panel assessed the validity of each HF diagnosis, ruling diagnoses as either objectified or non-objectified. Patients could be included in the study if they were 40 years or older and were registered in the participating practices.

Data sources

We used GPs' own EHR as a data source. Participating GPs used CareConnect, Medidoc (Corilus, Aalter), or HealthOne (HDMP, Anderlecht). These are three frequently used Belgian EHR software packages conforming to federal guidelines regarding structured data collection.¹² They operate on a MySQL relational database management system.¹³

Outcome

In the analysis of the classification sequence, we looked at GPs' assessment of HF before (coded or free-text HF diagnoses, i.e. registered HF) and after the audit, as well as the panel validation of post-audit HF diagnoses.

In our optimized audit model, we used the post-audit HF assessment by the GP as a binary outcome with the different queries of the extended audit as predictor variables.

Table 1 Trade-off between true positive rate or sensitivity and false positive rate for four different modelling approaches including bootstrapped confidence intervals

Model	AUC [95% CI]	True positive rate (sensitivity)	True negative rate (specificity) [95% CI]	False positive rate [95% CI]
GBM model	0.70 [0.65–0.77]	0.50	0.79 [0.68–0.85]	0.21 [0.15–0.33]
		0.75	0.56 [0.38–0.67]	0.44 [0.33–0.62]
		0.90	0.27 [0.18–0.38]	0.73 [0.62–0.82]
Logistic regression model	0.69 [0.64–0.75]	0.50	0.80 [0.71–0.85]	0.20 [0.15–0.29]
		0.75	0.48 [0.41–0.59]	0.52 [0.41–0.59]
		0.90	0.25 [0.16–0.41]	0.75 [0.59–0.84]
Simplified GBM model	0.70 [0.65–0.76]	0.50	0.80 [0.75–0.86]	0.20 [0.14–0.25]
		0.75	0.53 [0.41–0.63]	0.47 [0.37–0.59]
		0.90	0.23 [0.16–0.39]	0.77 [0.61–0.84]
Simplified logistic regression model	0.70 [0.65–0.75]	0.50	0.81 [0.75–0.86]	0.19 [0.14–0.25]
		0.75	0.53 [0.40–0.63]	0.47 [0.37–0.60]
		0.90	0.23 [0.16–0.39]	0.77 [0.61–0.84]

AUC, area under the curve; CI, confidence interval; GBM, gradient boosted machine.

Statistical analysis

We divided the total population at-risk for HF identified by the audits into two cohorts: patients who were identified as having HF by their GP after the audit (i.e. the OSCAR-HF study population) and everyone else. We described the proportion of positive responses on each query for both populations. In addition, we analysed audit performance using a logistic regression and gradient boosted machine (GBM) classification tree algorithm. A GBM algorithm sequentially combines many weak learning trees fitting upon the residuals of the previous one until finally achieving a strong learner.¹⁴ We calculated the relative importance of each predictor variable in our GBM model using Breiman's approach,¹⁵ based on the empirical improvement in squared error for each splitting variable, summed up over each boosting iteration. For the logistic model, we calculated relative importance based on the absolute value of the *t*-statistic for each model parameter. We consequently compared four models: a GBM model and logistic regression model utilizing all available predictor variables, and simplified GBM and logistic regression models based on the 10 most important predictor variables identified in the full GBM model. We designed and tested our models using a conventional training and test set partition correcting for class imbalance in our outcome and with a five-fold cross-validation. We plotted model performance on a receiver-operating characteristic (ROC) curve. We calculated optimal thresholds using Youden's index weighing the cost of a false negative classification as 10 times the cost of a false positive classification. We calculated area under the curve and specificities for different sensitivity cut-offs using bootstrapped confidence intervals. We additionally compared the trade-off between true positive and false positive cases in the audit in a precision–recall curve based on consecutive OR addition of the 10 most important queries. We calculated a number needed to screen for each query combination as the total number of non-HF patients each GP needs to screen for every 'true' HF patient. All statistical analyses were conducted in R Version 3.6.3 using R's 'caret' and 'pROC' packages.^{16,17} We included an additional analysis with panel objectified HF as outcome in Supporting Information, *Figure S1*.

Code mapping and model building

Supporting Information, *Table S1* provides an overview of the different queries and the coding of free-text syntax that were used to identify both registered and unregistered HF patients.

Medication information code mapping

We mapped all medications to underlying anatomical therapeutic chemical (ATC) classification codes and used seven different queries based on known HF medication classes or combinations.

Heart failure relevant well-known risk factors mapping

We mapped six different cardiovascular diseases to underlying International Classification of Primary Care (ICPC-2) codes or free-text constructs.

Heart failure relevant signs and symptoms mapping

We mapped four HF-specific signs and symptoms to underlying free-text strings. We mapped free-text decompensation as a separate query for one participating software package¹⁸ due to an idiosyncrasy in the underlying database architecture in order to assess the impact on overall audit performance.

Ethics

The OSCAR-HF pilot study conformed to the principles outlined in the Declaration of Helsinki. Before the study began, all participating GPs provided informed consent. An opt-out procedure was used for the identification of HF patients. Ethics committee approval was obtained from the University Hospitals Leuven Ethics Committee in November 2016 (B322201630391).

Results

Patient classification

Of the 18 011 patients eligible for the audit intervention, 4678 (26.0%) were identified as possible HF patients by the combined basic and extended audits and submitted for GPs' assessment in the audit stage (*Figure 1*). There were 310 patients with registered HF before GP assessment (basic audit),¹⁰ of whom 146 (47.1%) were judged not to have HF by their GP (over-registration). There were 538 patients with registered HF after GP assessment,¹⁰ of whom 374 (69.5%) did not have registered HF before GP assessment (under-registration).

Figure 2 further illustrates this classification flow for all patients with a GP assessment of before or after the audit ($n = 684$). There was a clear proportional difference in the panel validation of HF diagnoses between GP classified patients with or without registered diagnosis before GP assessment (84.8% objectified HF diagnoses vs. 58.6%, respectively).

Search strings

Figure 3 shows the 26 different queries of the extended audit in ascending order of frequency for the group with ($n = 538$) and without ($n = 4140$) a GP HF assessment after the audit. The three queries occurring most frequently in the audit pop-

Figure 1 Study flowchart describing the audit process in four stages. (1) Identification of the patient population aged 40 years or older. (2) Identification of possible heart failure (HF) patients through an extended audit in the electronic health record. (3) Identification of patients with registered HF before GP assessment. (4) GP assessment of the list generated in Step 2 as HF/no HF. GP, general practitioner.

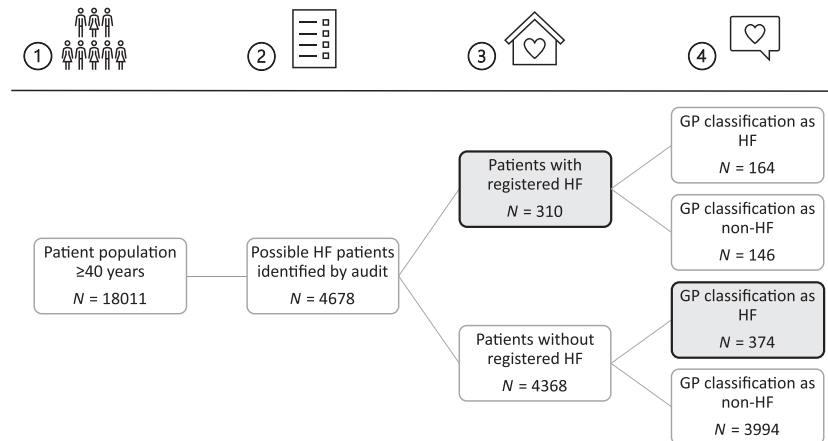
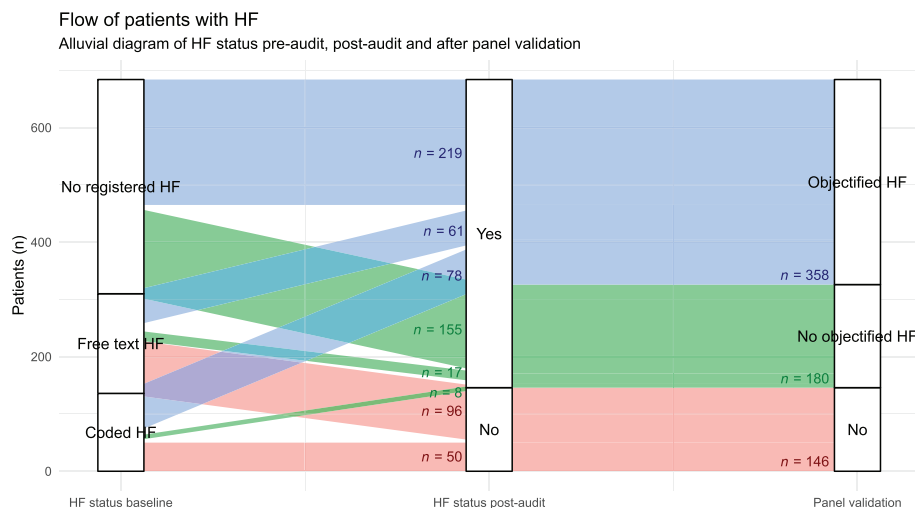


Figure 2 Alluvial diagram of the classification flow of patients who had registered HF before or after general practitioner assessment. HF, heart failure.



ulation were medication combinations (diuretics, beta-blockers, and angiotensin-converting enzyme inhibitors).

Prediction modelling

The variable importance for the logistic regression and gradient boosted models is depicted in *Figure 4*. The most important variables in each model were free-text cardiomyopathy and ischaemic heart disease, coded atrial fibrillation and mineralocorticoid receptor antagonists, and a combination of diuretics and angiotensin II receptor antagonists. *Figure 5* shows ROC curves for the four different models, as well as optimal thresholds using a weighted Youden's index with

corresponding sensitivities and specificities. *Table 1* shows the trade-off between increasing sensitivity and false positive rates for three pre-defined sensitivities (0.50, 0.75, and 0.90). The four models had comparable predictive performance as measured by the area under the curve.

Finally, *Table 2* and *Figure 6* express the trade-off between precision and recall (in this context also referred to as true positive rate and positive predictive value) in classification when we consecutively add the 10 most important query parameters as dichotomous OR operators in the audit population.

Using only four queries in an OR combination (free-text cardiomyopathy, ischaemic heart disease, mineralocorticoid receptor antagonists, and coded atrial fibrillation) was

Figure 3 Bar chart illustrating the frequency of positive responses for each query for the entire population identified in the audit ($n = 4679$). Green = patients assessed as HF by their GP, orange = patients assessed as non-HF by their GP. ACEi, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; BB, beta-blocker; C, coded query; FT, free-text query; GP, general practitioner; HF, heart failure.

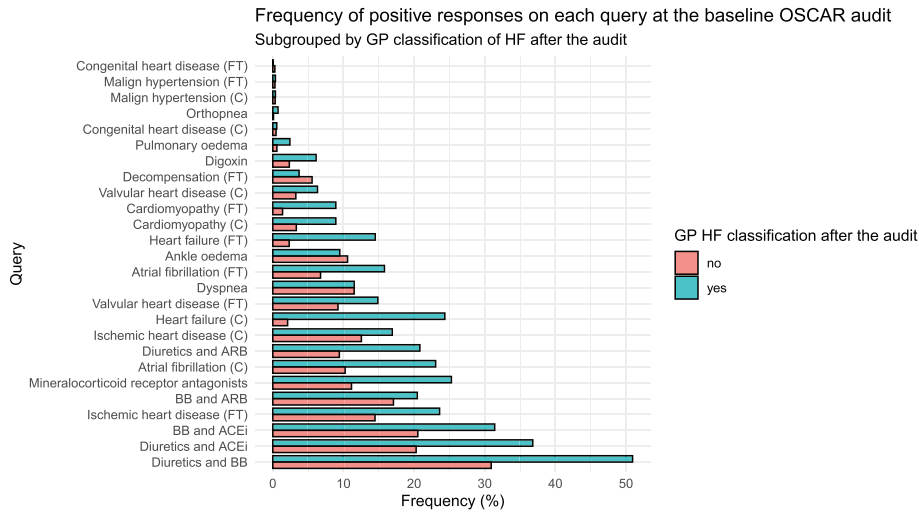


Figure 4 Variable importance score for each query for two different modelling strategies, scaled on the most important variable. ACEi, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; BB, beta-blocker; C, coded query; FT, free-text query; GBM, gradient boosted machine; GP, general practitioner; HF, heart failure.

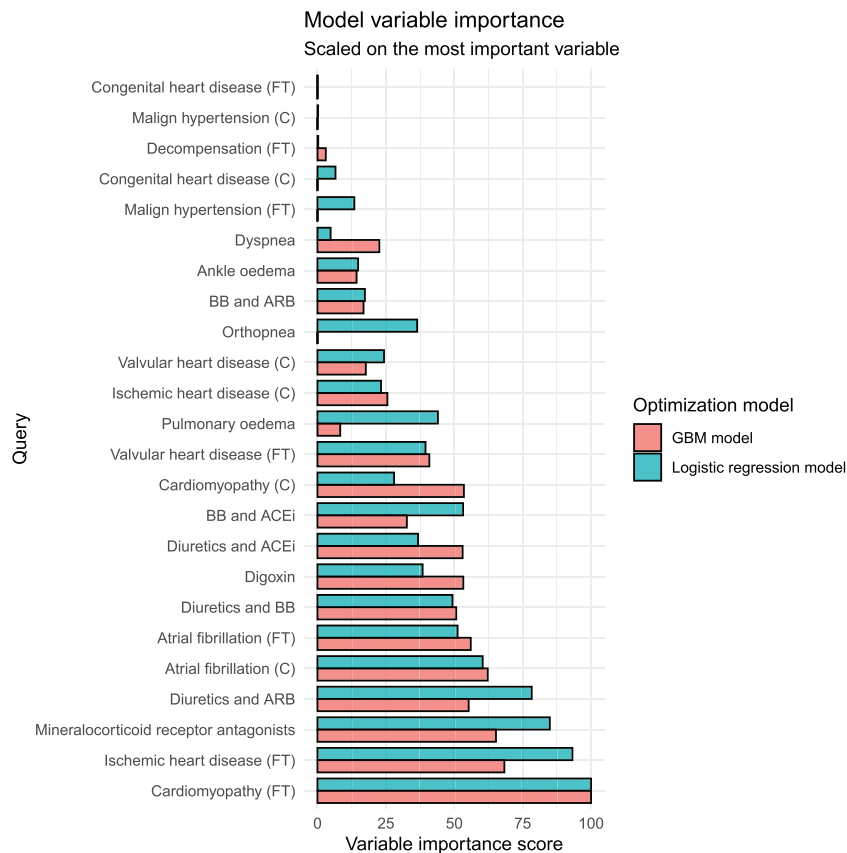


Figure 5 Comparison of sensitivity and specificity in a receiver operating characteristic curve for the identification of heart failure patients using four different modelling strategies. The labels indicate the sensitivity and specificity at the optimal cut-off for each strategy. GBM, gradient boosted machine.

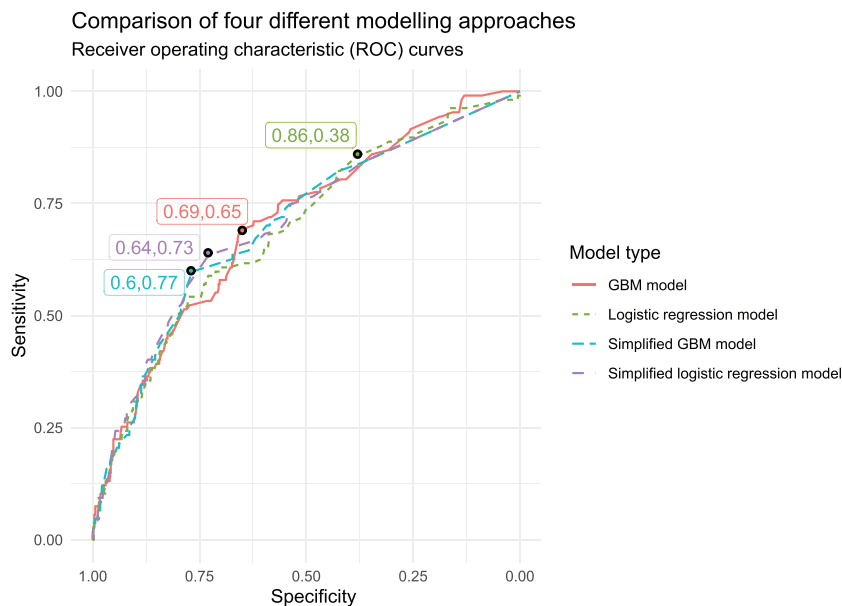


Table 2 Comparison of the positive predictive value and true positive rate in the electronic health record audit for a consecutive addition of the 10 most important dichotomous queries (gradient boosted machine model) in the electronic health record

Query	GP non-HF classification (n = 4340)	GP HF classification (n = 538)	Positive predictive value (precision)	True positive rate (recall)	Number needed to screen
CMP (FT)	57 (1.3%)	48 (8.9%)	0.46	0.09	1.19
+ IHD (FT)	649 (15.0%)	158 (29.4%)	0.20	0.29	4.11
+ MRA	1057 (24.4%)	252 (46.8%)	0.19	0.47	4.19
+ AF (C)	1383 (31.9%)	316 (58.7%)	0.19	0.59	4.38
+ AF (FT)	1465 (33.8%)	330 (61.3%)	0.18	0.61	4.44
+ Diuretics and ARB	1704 (39.3%)	369 (68.6%)	0.18	0.69	4.62
+ CMP (C)	1787 (41.2%)	382 (71.0%)	0.18	0.71	4.68
+ Digoxin	1816 (41.8%)	386 (71.7%)	0.18	0.72	4.70
+ Diuretics and ACEi	2242 (51.7%)	432 (80.3%)	0.16	0.80	5.19
+ Diuretics and BB	2680 (61.8%)	461 (85.7%)	0.15	0.86	5.81

ACEi, angiotensin-converting enzyme inhibitors; AF, atrial fibrillation; ARB, angiotensin receptor inhibitors; BB, beta-blockers; C, coded query; CMP, cardiomyopathy; FT, free-text query; GP, general practitioner; HF, heart failure; IHD, ischaemic heart disease; MRA, mineralocorticoid receptor antagonists.

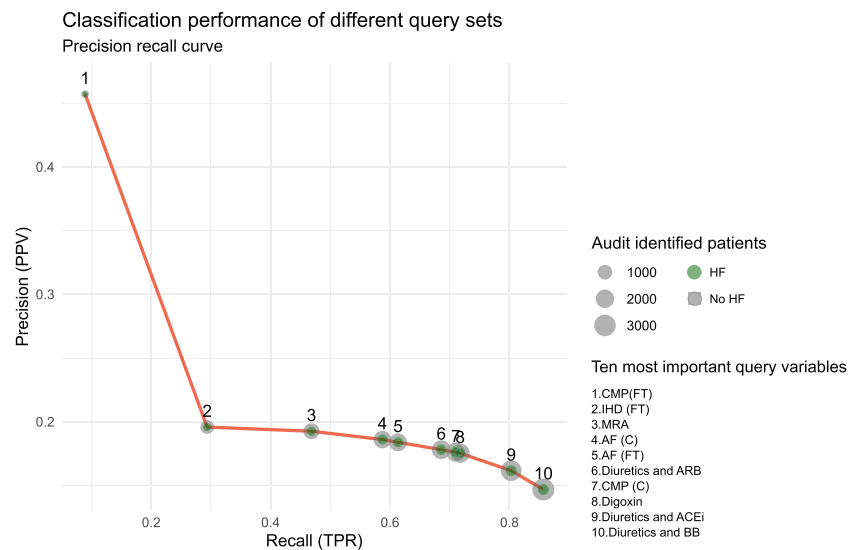
enough to identify almost 60% of GPs' self-assessed HF population, with GPs having to screen slightly more than four non-HF patients for every HF patient. Adding six additional queries increased identification to 86% but increased GPs' audit burden to six non-HF patients for every HF patient, which still amounted to a screening caseload reduction of 33% ($n = 1537$).

Discussion

Our study in 18 011 patients (recruited from 51 GPs) demonstrated that GPs' diagnostic coding accuracy of HF is

inadequate. Almost half of all patients with a registered HF diagnosis pre-audit were judged not to have HF following an audit in the EHR. Conversely, more than two-thirds of patients identified as having HF by their GP after the audit did not have a registered HF diagnosis before GP assessment. In addition, we demonstrated comparable performances on HF classification for both a classical statistical and machine learning approach. A simple query set using only search strings for cardiomyopathy, ischaemic heart disease, atrial fibrillation, digoxin, mineralocorticoid receptor antagonists, and combinations of renin-angiotensin system (RAS) inhibitors and beta-blockers with diuretics allowed identification of 86% of GPs' self-assessed HF population and a 33% reduction in screening caseload for the GP.

Figure 6 Precision–recall curve illustrating the benefit of adding queries with a logical OR operator in the entire data set. The y-axis indicates precision or positive predicative value (PPV). The x-axis indicates recall or true positive ratio (TPR). Integers in the graph depict the number of queries combined. The size of the circles expresses the total number of identified patients; the green inner circle expresses the proportion of OSCAR-HF patients. ACEi, angiotensin-converting enzyme inhibitors; AF, atrial fibrillation; ARB, angiotensin receptor inhibitors; BB, beta-blockers; C, coded query; CMP, cardiomyopathy; FT, free-text query; HF, heart failure; IHD, ischaemic heart disease; MRA, mineralocorticoid receptor antagonists.



With regard to our first aim, a description of the accuracy of registered HF in primary care records, the problem of EHR HF misclassification echoed previous studies demonstrating overdiagnosis of one-third to half of all HF patients^{4,5,19} and underdiagnosis of around one-sixth of patients in primary care settings, although this last finding was in a cohort of patients 65 or older presenting with dyspnoea on exertion.³ This is likely due to the difficulty of diagnosing HF in mainly older primary care HF populations and contrasts with the generally high validity of HF diagnoses in hospital records.^{20–22} In addition, we cannot rule out that several free-text HF diagnoses were working hypotheses that were later retracted by the GP, thus falsely elevating the level of registered HF. These findings have important consequences for health professionals trying to establish HF quality improvement (QI) initiatives in primary care because the clinical audits essential to the QI process²³ hinge upon a correct identification of target populations. Our findings indicate that this identification process cannot be solely contingent on elementary coded or free-text HF diagnoses and should first improve coding accuracy, for example, through an extended EHR audit with GP assessment. Interestingly, only a small number of patients ($n = 25$) in which GPs confirmed a pre-audit diagnosis of HF turned out to have non-objectified HF according to the experts. This suggests that performing even a very basic audit and feedback querying only the registered HF population can greatly enhance the specificity of registered HF in the primary care EHR.

Our EHR audit, composed of several coded and free-text search strings in a relational database management system,

used several identical coded comorbidities and medication combinations as previous primary care audits^{24–27} but also included free-text strings, signs and symptoms, and a larger variety of medication classes. Although algorithmic HF detection through administrative data sources has been extensively studied, most studies use coded information in hospital records,^{28,29} limiting validity in outpatient settings. This is the first study analysing such an EHR algorithm in primary care records. Our findings of a clear improvement in HF identification with diminishing marginal returns for increased query sets and the identification of those queries that are most important to HF identification in this primary care population are therefore particularly relevant. The adequate registration of these queries could be the focus of QI initiatives such as EHR training for GPs or natural language processing tools performing back-end query registration in the EHR, although it remains to be established whether more accurate registration leads to improved health outcomes and whether the substantial upfront time investment incumbent on even an optimized search algorithm limits implementation in routine clinical practice.

Strengths and limitations

The major strengths of our study are the extensive diagnostic validation process involving patients' own GP and a decision tree approach mirroring real-world binary query combinations in the EHR with logical AND/OR operators. There are a few limitations, however. First, we used GPs'

assessment of HF as the outcome of interest for our classification models, rather than objectified HF according to the expert panel. This could lead to an overestimation of the true HF population. However, this reflects real-world implementation, where GPs' judgement on HF would be the outcome rather than an expert panel HF diagnosis. Moreover, this audit should be regarded as part of a continuous QI process rather than a precise diagnostic instrument. For example, a GP could be triggered by the audit to refer patients to a cardiologist for validation when uncertain about an HF diagnosis, ideally leading to rapid feedback on his diagnostic decision-making. In addition, our study design was not ideal for objectively quantifying the accuracy of HF diagnoses in unaudited GP HF records, because patients with a registered HF diagnosis and a 'non-HF' assessment by their GP were not presented for panel validation. Second, we were somewhat confined in the audit optimization process because we could only test variables on patients identified in the extended audit procedure, thus potentially missing patients in the broader primary care population of >18 000 patients. However, the extended audit was comprehensive and conducted in primary care practices with good EHR registration. Because HF prevalence levels in the extended audit population were in accordance with other primary care studies,¹⁰ this 'dark number' is likely low. Third, we did not use other potentially more powerful machine learning models, such as neural networks or more recent decision tree algorithms, which could have negatively impacted classification performance. However, we chose the classical GBM model because of its well-established method to assess our primary objective, variable importance.

References

- Conrad N, Judge A, Tran J, Mohseni H, Hedgecote D, Crespillo AP, Allison M, Hemingway H, Cleland JG, McMurray JJV, Rahimi K. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *Lancet (London, England)* 2018; **391**: 572–580.
- Groenewegen A, Rutten FH. Near-home heart failure care. *Eur J Heart Fail* 2019; **21**: 110–111.
- van Riet EE, Hoes AW, Limburg A, Landman MA, van der Hoeven H, Rutten FH. Prevalence of unrecognized heart failure in older persons with shortness of breath on exertion. *Eur J Heart Fail* 2014; **16**: 772–777.
- Dahlstrom U, Hakansson J, Swedberg K, Waldenstrom A. Adequacy of diagnosis and treatment of chronic heart failure in primary health care in Sweden. *Eur J Heart Fail* 2009; **11**: 92–98.
- Valk MJ, Mosterd A, Broekhuizen BD, Zuithoff NP, Landman MA, Hoes AW, Rutten FH. Overdiagnosis of heart failure in primary care: a cross-sectional study. *Br J Gen Pract* 2016; **66**: e587–e592.
- Hull SA, Rajabzadeh V, Thomas N, Hoong S, Dreyer G, Rainey H, Ashman N. Improving coding and primary care management for patients with chronic kidney disease: an observational controlled study in East London. *Br J Gen Pract* 2019; **69**: e454–e461.
- Nitsch D, Caplin B, Hull S, Wheeler D, Kim L, Cleary F. National chronic kidney disease audit: national report (Part 1). Healthcare Quality Improvement Partnership. 2017.
- Hartung DM, Hunt J, Siemieniczuk J, Miller H, Touchette DR. Clinical implications of an accurate problem list on heart failure treatment. *J Gen Intern Med* 2005; **20**: 143–147.
- Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, O'Brien MA, Johansen M, Grimshaw J, Oxman AD, Cochrane Effective Practice and Organisation of Care Group. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev* 2012: CD000259.
- Smeets M, Vaes B, Aertgeerts B, Raat W, Penders J, Vercammen J, Droogne W, Mullens W, Janssens S. Impact of an extended audit on identifying heart failure patients in general practice: baseline results of the OSCAR-HF pilot study. *ESC Heart Fail* 2020; **7**: 3950–3961.
- Smeets M, Aertgeerts B, Mullens W, Penders J, Vercammen J, Janssens S, Vaes B. Optimising standards of care of heart failure in general practice the

Conclusions

Diagnostic coding of HF in primary care health records is inaccurate with a high degree of underclassification and overclassification. An optimized query set using only search strings for cardiomyopathy, ischaemic heart disease, atrial fibrillation, digoxin, mineralocorticoid receptor antagonists, and combinations of RAS inhibitors and beta-blockers with diuretics enabled identification of more than 80% of GPs' self-assessed HF population, albeit with modest specificity.

Conflict of interest

W.R., M.S., S.H., B.A., J.P., W.M., and B.V. report no conflict of interest. S.J. is holder of a named chair in Cardiology at the University of Leuven financed by Astra-Zeneca.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Overview of queries used to identify registered HF patients as well as patients at risk for HF in the electronic health record. ICPC-2 = International Classification of Primary Care Second edition.

Figure S1. Frequency of positive responses on each query for OSCAR cohort, subgrouped by panel classification of HF after the audit.

- OSCAR-HF pilot study protocol. *Acta Cardiol* 2018; **74**: 371–379.
12. RAMIT. In Volksgezondheid F., ed. *Registration of Electronic Health Records for Use in General Practice*. Belgium; 2019.
 13. MySQL 8.0 Reference Manual—What is MySQL?: Oracle; 2020. Available from: <https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html>. Accessed 17 Dec 2020.
 14. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York: Springer; 2013.
 15. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001; **29**: 1189–1232.
 16. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008; **28**: 26.
 17. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform* 2011; **12**: 77.
 18. Medidoc. Aalter, Belgium: Corilus.
 19. Verdu-Rotellar JM, Frigola-Capell E, Alvarez-Perez R, da Silva D, Enjuanes C, Domingo M, Mena A, Munoz MA. Validation of heart failure diagnosis registered in primary care records in two primary care centres in Barcelona (Spain) and factors related. A cross-sectional study. *Eur J Gen Pract* 2017; **23**: 107–113.
 20. Cozzolino F, Montedori A, Abraha I, Eusebi P, Grisci C, Heymann AJ, Lombardo G, Mengoni A, Orso M, Ambrosio G. A diagnostic accuracy study validating cardiovascular ICD-9-CM codes in healthcare administrative databases. The Umbria Data-Value Project. *PLoS ONE* 2019; **14**: e0218919.
 21. Schaufelberger M, Ekestubbe S, Hultgren S, Persson H, Reimstad A, Schaufelberger M, Rosengren A. Validity of heart failure diagnoses made in 2000–2012 in western Sweden. *ESC Heart Fail* 2020; **7**: 36–45.
 22. 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**: 787–791.
 23. Backhouse A, Ogunlayi F. Quality improvement into practice. *BMJ* 2020; **368**: m865.
 24. Frigola-Capell E, Verdu-Rotellar JM, Comin-Colet J, Davins-Miralles J, Hermsilla E, Wensing M, Suñol R. Prescription in patients with chronic heart failure and multimorbidity attended in primary care. *Qual Prim Care* 2013; **21**: 211–219.
 25. Cancian M, Battaggia A, Celebrano M, Del Zotti F, Novelletto BF, Michieli R, Saugo M, Pellizzari M, Toffanin R. The care for chronic heart failure by general practitioners. Results from a clinical audit in Italy. *Eur J Gen Pract*; **19**: 3–10.
 26. 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**: 160–166.
 27. Gini R, Schuemie MJ, Mazzaglia G, Lapi F, Francesconi P, Pasqua A, Bianchini E, Montalbano C, Roberto G, Barletta V, Cricelli I, Cricelli C, Dal Co G, Bellentani M, Sturkenboom M, Klazinga N. Automatic identification of type 2 diabetes, hypertension, ischaemic heart disease, heart failure and their levels of severity from Italian general practitioners' electronic medical records: a validation study. *BMJ Open* 2016; **6**: e012413.
 28. 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**: 129–140.
 29. 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**: e104519.