

Cardiology Hospital Admission Risk Prediction (CHARP). Training, internal validation and technical implementation in the Electronic Health Record (EHR).

Selder JL¹, Witteman OV¹, Meer van der O², Boonstra MJ¹, Zwaard van de S¹, Asselbergs
FW^{1,3,4}, Schijven M^{5,6,7}, Rossum AC¹, Allaart CP¹

Institutions

1: Department of Cardiology, Amsterdam Cardiovascular Sciences, Amsterdam University
Medical Centre, University of Amsterdam, Amsterdam, The Netherlands

2: Department of ICT, Amsterdam University Medical Centre, University of Amsterdam,
Amsterdam, The Netherlands

3: The National Institute for Health Research University College London Hospitals
Biomedical Research Centre, University College London, London, United Kingdom

4: Institute of Health Informatics, University College London, London, United Kingdom

5: Department of Surgery, Amsterdam UMC, Location University of Amsterdam,
Amsterdam, the Netherlands

6: Amsterdam Gastroenterology and Metabolism, Amsterdam UMC, University of
Amsterdam, The Netherlands

7: Amsterdam Public Health, Digital Health, Amsterdam UMC, Amsterdam, the
Netherlands

1 Corresponding author

2 Selder JL, j.selder@amsterdamumc.nl, +31645256921

4 Abstract

5 Background

6 Rising healthcare demand is increasingly outpacing available outpatient
7 capacity in cardiology, where follow-up is often scheduled at fixed
8 intervals despite substantial variation in individual patient risk. This
9 uniform follow-up approach contributes to high outpatient workload
10 and inefficient use of clinical resources. Accurate risk estimation using
11 routinely collected electronic health record (EHR) data may support
12 more individualized follow-up planning by identifying patients at very
13 low risk of mortality or unplanned hospitalization, in whom follow-up
14 intervals could be safely extended.

15 Methods

16 We developed and validated a machine-learning model as part of the
17 Cardiology Hospital Admission Risk Prediction (CHARP) program. The

retrospective baseline cohort comprised 307,792 outpatient visits from 52,989 unique patients at Amsterdam UMC. The primary endpoint was a composite of unplanned cardiac hospitalization or all-cause death within 2 years; the 1-year composite endpoint served as a secondary outcome. Model development and validation were performed in a filtered, leakage-safe prediction cohort using gradient-boosted decision trees (XGBoost) with strict patient-level GroupKFold cross-validation. All predictions and performance metrics were retrospectively evaluated at the outpatient visit (trigger) level. Retrospective model performance was assessed using AUROC, AUPRC, Brier score, calibration curves, and SHAP-based explainability. The final model was technically deployed within the electronic health record to allow automated, visit-level risk estimation in a prospective silent-running environment.

Results

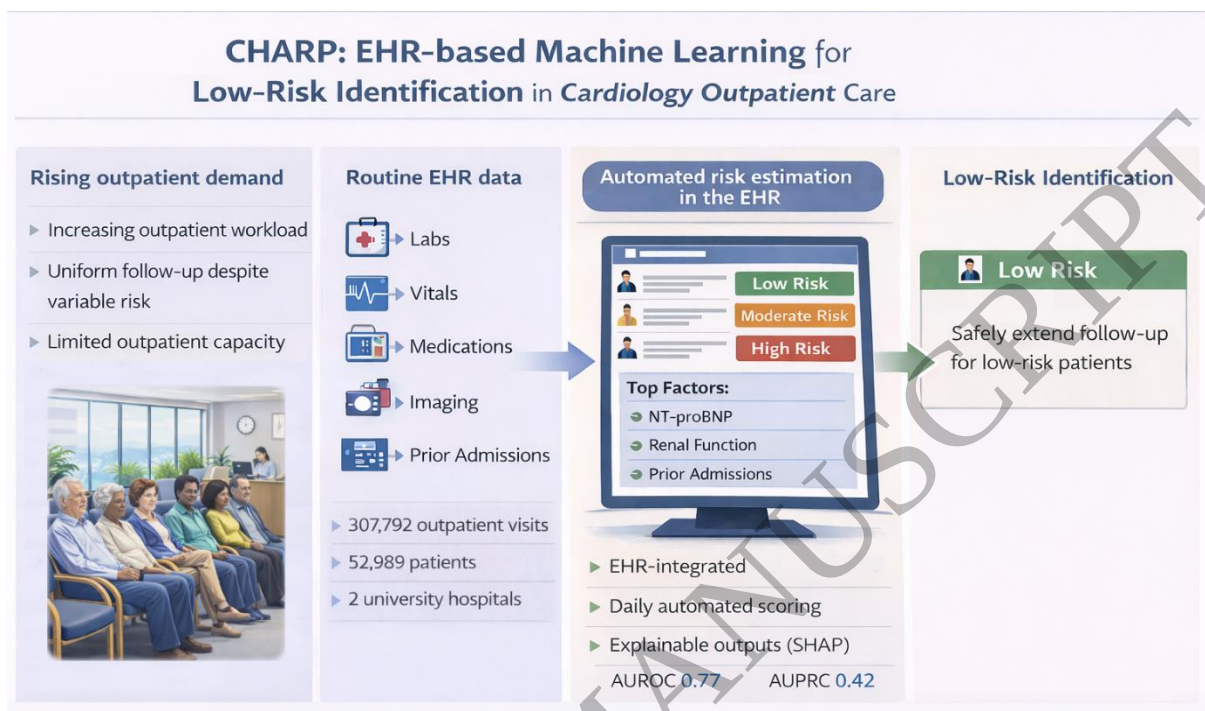
In the filtered prediction cohort (199,961 visits), the 2-year composite endpoint prevalence was 16.8%. Across five cross-validation folds, the CHARP model achieved a mean AUROC of 0.77 ± 0.00 and AUPRC of

0.42 $\pm$ 0.01, with a Brier score of 0.12, indicating strong overall discrimination and good calibration. Key predictors included NT-proBNP, renal function indices, prior hospitalizations, and cardiac function measures. The deployed CHARP pipeline successfully generated daily risk predictions for all scheduled cardiology outpatients in the EHR environment throughout the silent-running period.

Interpretation

This study shows that machine-learning applied to routine EHR data can deliver clinically meaningful, visit-level risk stratification for cardiology outpatients. The successful EHR integration of CHARP enables prospective evaluation of data-driven follow-up strategies aimed at reducing outpatient clinic burden through safe de-intensification of follow-up for low-risk patients.

1 Graphical Abstract



2

3 Keywords

- 4 Artificial intelligence (AI)
- 5 Electronic health record
- 6 Machine learning
- 7 Hospital admission risk prediction
- 8 Cardiology outpatients
- 9 Clinical decision support

1 Introduction

2 Healthcare demand in Europe is projected to rise substantially over the
3 next 15 years, driven by population growth, aging societies, and
4 changing disease patterns, according to the World Health Organization
5 (WHO).(1) Although the number of physicians and nurses has increased
6 by approximately 10% over the past decade, this growth is unlikely to
7 keep pace with future demand.(1) Even if workforce targets are to be
8 met, this would be costly, as wages and benefits account for
9 approximately 60% of total hospital expenditure.(2) This makes it
10 imperative to explore innovative strategies to meet rising healthcare
11 needs while using resources efficiently.

12 A promising approach is the use of Artificial Intelligence (AI) for clinical
13 risk prediction, particularly for unplanned hospital admissions.(3)

14 Accurate prediction of patients' future risk may enable clinicians to
15 identify patients at very low risk in whom follow-up intervals can be
16 safely extended or routine outpatient visits reduced, thereby relieving
17 pressure on outpatient capacity. Such an approach focuses on

1 supporting safe de-intensification of routine follow-up while
2 maintaining existing care pathways and clinical oversight. In this way,
3 risk-informed follow-up planning may contribute to more efficient use
4 of healthcare resources without increasing clinical workload.

5 Although several studies have applied machine learning (ML) to
6 admission or readmission risk prediction, reported performance varies
7 widely and is often difficult to compare due to differences in study
8 populations, outcome definitions, data availability, and methodological
9 limitations such as lack of external validation or potential data
10 leakage.(4–19) Reported areas under the receiver operating
11 characteristic curve (AUROCs) range from 0.65 to 0.94 across a variety
12 of algorithms, including logistic regression, random forests, gradient
13 boosting, support vector machines, and deep neural networks. While
14 gradient boosting methods frequently achieve strong performance on
15 structured electronic health record (EHR) data, model performance is
16 highly dependent on local care processes, patient case-mix, and data

1 structure. Consequently, transferability across clinical settings is often
2 limited, and models developed in one healthcare system may not
3 generalize to others without local development and evaluation. In
4 particular, machine learning–based prediction of hospital admission risk
5 has not been evaluated in a heterogeneous Dutch tertiary-care
6 cardiology population using routinely collected EHR data within an
7 implementation-oriented framework.

8 Aim of the Study

9 In the context of an anticipated mismatch between healthcare demand
10 and available workforce, the overarching program aims to improve the
11 efficiency of cardiology outpatient care through data-driven decision
12 support by identifying patients at very low risk of hospital admission,
13 thereby enabling reduction of unnecessary routine follow-up while
14 maintaining patient safety within existing care pathways. Within this
15 program, the present study specifically aimed to develop, train, and
16 internally validate a machine-learning model (CHARP) to estimate the
17 risk of a composite endpoint of unplanned cardiac hospitalization or all-

1 cause death within 2 years following an outpatient cardiology visit.
2 Secondary objectives were to assess model performance for the same
3 composite endpoint at 1 year and to deploy the algorithm within the
4 hospital EHR to enable silent-running and real-time generation of visit-
5 level risk estimates.

7 Methods

8 Dataset and Data Sources

9 A pseudonymized dataset was constructed from cardiology outpatients
10 of Amsterdam UMC (locations Amsterdam Medical Center (AMC) and
11 Vrije Universiteit University Medical Center (VUMC)). Patients with at
12 least one outpatient cardiology visit between January 2016 and June
13 2025 were included. Data extraction and harmonization were
14 performed in Snowflake (Snowflake Inc., Bozeman, MT, USA) using
15 curated dynamic views. Clinical variables were standardized across Epic
16 systems using LOINC and SNOMED CT terminology by Medscio

(Medscio BV, Amsterdam, The Netherlands). The dataset comprised demographics (age, sex), vital signs, laboratory results, medication exposure, outpatient and inpatient encounters, Diagnosis Treatment Combinations (DBC's), procedures and implantable devices, echocardiography-derived left ventricular ejection fraction, and imaging metadata (modality and timestamp only; no pixel data). Operational definitions of key clinical variables as implemented in Snowflake views are provided in Appendix B. Each row represented a trigger event, defined as an individual outpatient cardiology visit, serving as the temporal anchor for feature derivation and outcome assessment. Predictors were selected a priori based on their availability at or before the outpatient visit, scalability for daily automated deployment in the Epic environment, and robustness against information leakage. We therefore prioritized routinely collected, structured EHR variables that are consistently recorded as discrete fields and harmonized across sites. Variables requiring limited additional processing but with high clinical relevance, such as left ventricular ejection fraction (LVEF), were

1 included using a dedicated extraction workflow. Unstructured data
2 sources (e.g., radiology narratives or free-text reports) were
3 intentionally excluded in this initial version as it would substantially
4 increase pipeline complexity, governance requirements, and leakage
5 risk in a real-time deployment setting.

6 Outcome Definition

7 The target variable (TARGET_EVENT) indicated whether unplanned
8 cardiac hospitalization or death occurred after the trigger event.

9 Primary endpoint: composite of unplanned cardiac hospitalization and
10 all-cause death within 2 years after the trigger date. Secondary
11 endpoint: same composite within 1 year.

12 Unit of Analysis

13 The unit of analysis was the trigger level, defined as an individual
14 outpatient cardiology visit. Patients could contribute multiple triggers
15 over time. All analyses were performed at the trigger level, with
16 performance metrics reflecting visit-level risk prediction rather than
17 patient-level outcomes. Trigger visits included all outpatient visits

1 involving a cardiologist or cardiology resident within the Amsterdam
2 UMC cardiology outpatient clinics (AMC and VUMC). This included a
3 broad spectrum of both general and subspecialty cardiology patients,
4 ranging from relatively common outpatient conditions such as coronary
5 artery disease, valvular disease, arrhythmias, and heart failure to more
6 specialized clinics. Visits conducted exclusively by heart failure nurse
7 specialists or atrial fibrillation nurse specialists were not included. Visits
8 outside the cardiology outpatient clinics were also not included.

9 Measures to prevent information leakage, including patient-level data
10 separation and handling of temporally clustered visits, are described in
11 detail in the *Leakage Prevention* section.

12 Hospitalization

13 Hospital admissions were defined as all unplanned admissions to one of
14 the following wards: AMC cardiology ward; AMC coronary care unit;
15 AMC acute cardiac emergency unit; VUMC coronary care unit; VUMC
16 acute cardiac emergency unit; VUMC cardiology ward; AMC
17 cardiothoracic surgery ward; AMC adult intensive care unit (ICU); and

VUMC adult intensive care unit (ICU). Admissions were classified as cardiac based on ward type and treating specialty, in accordance with institutional clinical practice. Unplanned admissions were prospectively recorded by nursing staff using a standardized admission-type field within the electronic health record, in which admissions were coded as planned or unplanned as part of routine clinical workflow.

Mortality

Deaths were identified through linkage with municipal registries (Statistics Netherlands, CBS) and were labeled as positive if all-cause mortality occurred within the predefined prediction horizon of 2 years (primary objective) and 1 year (secondary objective).

Leakage Prevention

To prevent information leakage and ensure unbiased performance estimation, multiple safeguards were implemented throughout data preprocessing and model development:

- Exclusion of near-event triggers: Outpatient visits occurring within 24 hours prior to an unplanned hospital admission were entirely

1 excluded, as these visits are likely to reflect imminent admission-
2 related diagnostics or administrative processes rather than true
3 outpatient decision points.

- 4 • Outcome masking: All outcome-related variables (e.g., EVENT_*
5 flags, admission indicators, mortality labels) were removed from
6 the feature set prior to model training.
- 7 • Strict temporal anchoring: All predictors were derived using only
8 information available at or before the trigger date. Raw
9 timestamp variables (*_DATE) were used solely for feature
10 derivation and subsequently removed from the final feature
11 matrix to prevent implicit temporal leakage.
- 12 • Follow-up requirement and outcome completeness: Triggers with
13 an observed event were retained regardless of follow-up
14 duration, as event occurrence within the 2-year prediction
15 horizon can be ascertained with certainty. For non-event triggers,
16 a minimum follow-up of 550 days was required. This threshold,

1 rather than the full 730 days, was chosen as a pragmatic
2 compromise to avoid unnecessary exclusion of recent,
3 predominantly event-free outpatient visits while limiting false-
4 negative outcome labeling due to incomplete follow-up.

5 Additional sensitivity analyses requiring full 730-day follow-up
6 were performed.

- 7 • Near-duplicate trigger deduplication: To reduce redundancy and
8 burst effects, near-duplicate outpatient triggers occurring within
9 90 days for the same patient were deduplicated in validation and
10 test sets.
- 11 • Patient-level data separation: All cross-validation and holdout
12 splits were performed at the patient level, ensuring that data

1 from the same patient never appeared simultaneously in training
2 and validation sets.

- 3 • Consistent preprocessing: All preprocessing steps were
4 implemented within a single version-controlled pipeline and
5 applied identically across all validation scenarios.

6 Feature Engineering

7 Feature derivation and engineering were implemented in a custom Python pipeline
8 (PipelineCHARPPlus).

- 9 • Healthcare and medication utilization:
 - 10 ○ UTIL_INDEX: aggregated exponential-decay index of prior encounters and
11 procedures, $w = e^{-\lambda \cdot \Delta t}$, $\lambda = 0.01$ per day.
 - 12 ○ UTIL_COUNT features: counts of outpatient visits, admissions, and
13 telephone contacts in the preceding 30, 90, and 180 days.
 - 14 ○ Total number of active diagnosis–treatment combinations (DBC), the Dutch
15 reimbursement unit representing an episode of specialist care for a specific
16 diagnosis, and the number of newly opened DBCs in the preceding 14
17 months.
 - 18 ○ Medication counts: number of concurrent medications at the trigger date
19 and cumulative prescriptions in the prior 14 months.
- 20 • Laboratory results:

- LV (last value) and BLV (before-last value) prior to the trigger date.
- Delta features ($\Delta = LV - BLV$).
- Min/max values for selected biomarkers (e.g., maximum NT-proBNP, maximum troponin).
- Missing values were kept unchanged, allowing the model to use whether or not a test was performed as potentially informative.
- Vital signs: last and before-last values for blood pressure, heart rate, and BMI, with corresponding delta features.
- Renal function: Chronic Kidney Disease (CKD) stage derived from most recent eGFR, with an additional time-weighted score (CKD_SCORE_TIMEWEIGHTED, $\lambda \approx 0.002/\text{day}$).
- Cardiac function: LVEF values were extracted using a dedicated semi-automated pipeline combining custom parsing of echocardiography PDF reports with structured retrieval from the GE EchoPAC database (EchoPAC, GE HealthCare, Chicago, Illinois, USA). Successfully extracted values were included in the modeling pipeline, while missing values were handled natively by XGBoost. Extraction accuracy was manually verified in a random subset of patients. Missing values largely reflected heterogeneous report structures in earlier years and the semi-automated nature of the extraction workflow, rather than absence of echocardiographic assessment itself. Procedures and imaging: counts of echocardiograms, angiographies, and other structured procedures.

Model Development

Gradient boosting with XGBoost (binary logistic objective) was used for risk prediction. XGBoost was selected because of its strong empirical performance on large structured EHR datasets, its ability to model nonlinear relationships and feature interactions, its native handling of missing data without requiring explicit imputation (although this does not resolve the reasons for missingness). All feature derivation and aggregation were completed upstream in Snowflake, after which the finalized feature matrix was exported for model training and evaluation in Python. Hyperparameters were optimized using Optuna, with the harmonic mean of AUROC and AUPRC as the optimization objective. A representative configuration included a learning rate of approximately 0.028, maximum tree depth of 10, minimum child weight of 6, subsampling rate of approximately 0.787, column subsampling rate of approximately 0.551, L2 regularization (λ) of approximately 0.492, L1 regularization (α) of approximately 8.541 and up to 1000 boosting

rounds with early stopping after 50 rounds. The overall modeling workflow is summarized in Appendix A.

Missing value handling

XGBoost's native missing value handling was used. During tree construction, missing values are automatically directed to the branch that optimizes the split criterion. No external imputation was performed.

Evaluation

Cross-validation: Model performance was evaluated using 5-fold GroupKFold cross-validation with strict patient-level separation. Out-of-fold predictions were used for performance assessment and subgroup analyses. Discrimination was quantified using AUROC and AUPRC (mean $\pm$ SD across folds), along with sensitivity, specificity, and precision across probability thresholds. Calibration was assessed on out-of-fold predictions using calibration plots and Brier score; isotonic regression

was applied for calibration assessment only and not used to modify model predictions.

Explainability: Model-level: XGBoost's native importance metrics, gain, weight, and total gain, were extracted across folds, averaged, and visualized for the top 30 predictors.

Trigger-level: SHAP (Shapley Additive Explanations) values were computed for each individual trigger to identify the most influential features contributing to the patient's risk score. In the EHR, the five most important SHAP-ranked features were displayed alongside the prediction.

Generalizability and Cross-site Validation

We evaluated cross-site model transportability using four leakage-free validation scenarios across the two Amsterdam UMC hospital sites (AMC and VUMC), which historically operated as semi-independent tertiary care environments with separate EHR systems for most of the study period. Two scenarios involved full site-exclusive validation,

1 where the model was trained on one site and tested on the other. In
2 addition, we performed patient-exclusive 15% internal holdout
3 validations for each site (ALL → VUMC and ALL → AMC), ensuring that
4 no patient contributed data to both training and testing. The same
5 leakage-prevention procedures used in the main pipeline, including
6 patient-level exclusivity and burst-deduplication of temporally
7 clustered encounters, were applied identically in all cross-site validation
8 scenarios. Although these analyses do not constitute formal external
9 validation, they provide an intermediate assessment of model
10 transportability across distinct clinical environments within a single
11 evolving healthcare system. Model discrimination was assessed using
12 AUROC and AUPRC.

13 Model Deployment

14 The trained model was serialized to JSON (JavaScript Object Notation)
15 and deployed in the Amsterdam UMC clinical infrastructure. Execution

occurs on a high-performance computing (HPC) cluster positioned between Snowflake (data warehouse) and the Epic EHR.

Each morning, the pipeline runs automatically on all cardiology outpatients scheduled for that day. For every patient, the system generates a probability of unplanned hospitalization or death within 2 years. Alongside the risk score, the five most influential features (SHAP-ranked at trigger level) are returned.

Outputs are integrated in the EHR in silent-running mode and are accessible for research purposes, without being used to guide or influence clinical decision-making. The retrospective analyses presented in this study did not include data from the prospective silent-running deployment. The silent-running setup is intended to facilitate future prospective monitoring of model performance and usability in routine clinical practice, providing the foundation for prospective validation and subsequent randomized evaluation.

Software and Environment

All analyses were conducted in Python (pandas, NumPy, scikit-learn, XGBoost, Optuna, Matplotlib, Seaborn) within Jupyter notebooks. Extensive data preprocessing was executed on Snowflake.

Risk of bias and applicability

Risk of bias and applicability were evaluated according to the PROBAST+AI framework, covering participant selection, predictors, outcome definition, analysis, and AI-specific leakage domains.

STROBE and CODE-EHR

This study was reported in accordance with the STROBE statement for cohort studies and the CODE-EHR framework for structured healthcare data. See Supplementary Data 1 and 2.

Results

Study characteristics

Baseline Characteristics

A total of 52,989 unique patients were included, contributing 307,792 outpatient cardiology visits that served as index triggers for risk prediction. Unless stated otherwise, baseline characteristics are reported at the outpatient visit (trigger) level. The median age at the time of visit was 63 years (interquartile range [IQR] 49–73), and 57% of visits involved male patients. The study population represents a broad outpatient cardiology cohort, with a median body mass index of 25.8 kg/m² (IQR 23.1–29.3), a median heart rate of 70 bpm (IQR 61–80), and a mean arterial pressure of 95 mmHg (IQR 87–104). Hypertension and ischemic heart disease were present in approximately 30% of visits, while 27% involved patients with a history of heart failure and 20% with diabetes mellitus. Renal dysfunction (eGFR <60 mL/min/1.73 m²) was observed in 21% of visits, and 3.5% in severe form (eGFR <30 mL/min/1.73 m²). Device therapy was common, with 13% of visits

1 involving patients with an implantable cardioverter-defibrillator (ICD)
2 and 2% a CRT-D system. Exact definitions and SQL logic used to derive
3 comorbidity and device flags are provided in Appendix B (Snowflake
4 definitions).

5 Laboratory and Imaging

6 At the outpatient visit, the median NT-proBNP was 406 ng/L (IQR 130–
7 1310), troponin T 16 ng/L (IQR 8–44), and eGFR 67 mL/min/1.73 m²
8 (IQR 59–86). Median left ventricular ejection fraction (LVEF) was 58%
9 (IQR 50–67), among visits with available LVEF data, 11.4% had an LVEF
10 ≤40%.

11 Healthcare Utilization and Medication

12 During the preceding 14 months, patients had a mean of 1.9 ± 1.8
13 hospitalizations and 7.7 ± 8.2 hospital contacts, including 3.7 ± 3.0
14 outpatient or telephone visits.

15 Use of medications commonly recommended in cardiovascular
16 guidelines was frequent: 44% of outpatient visits involved β -blockers,
17 36% ACE inhibitors/ARBs/ARNIs, 16% mineralocorticoid receptor

1 antagonists, and 26% loop diuretics; 3.5% involved SGLT2 inhibitors.

2 Medication exposure was determined through structured medication
3 records up to the outpatient visit, based on substance-level mappings
4 and ATC (Anatomical Therapeutic Chemical)-class groupings.

5 Event Rates

6 Within 2 years after the outpatient visit, the composite endpoint of
7 unplanned cardiac hospitalization or death occurred in 19.1% of visits
8 (14.4% hospitalization, 6.4% death). At 1 year, the composite endpoint
9 occurred in 12.5% (9.9% hospitalization, 3.5% death). These event rates
10 reflect the unfiltered baseline cohort prior to application of model-
11 specific preprocessing and leakage-prevention steps, and are
12 summarized in Table 1.

13

Variable	Overall	2y Event: Yes	2y Event: No	Per patient: first visit	Per patient: last visit	Missi ng %	Unit / Definition
N unique patients	52989					0.0	Distinct pts
N outpatient visits	307792					0.0	All rows
Baseline characteristics							
Age	63.0 [49.0–73.0]	69.0 [57.0–78.0]	61.0 [47.0–72.0]	59.0 [44.0–71.0]	62.0 [46.0–74.0]	0.0	years
Sex: male	175440 (57.0%)	33979 (58.5%)	141461 (56.7%)	29224 (55.2%)	29224 (55.2%)	0.0	n (%)
Body mass index	25.8 [23.1–29.3]	26.1 [23.3–29.8]	25.8 [23.1–29.2]	25.5 [22.8–28.9]	25.5 [22.8–28.9]	17.1	kg/m ²
Heart rate	70.0 [61.0–80.0]	71.0 [62.0–81.0]	70.0 [61.0–80.0]	71.0 [62.0–82.0]	71.0 [62.0–81.0]	12.8	bpm
Mean arterial pressure	95.3 [86.7–104.3]	93.7 [84.3–103.0]	95.7 [87.3–104.7]	95.3 [86.7–104.7]	95.7 [87.0–104.7]	12.3	mmHg
Key risk factors							
Hypertension label	92426 (30.0%)	23413 (40.3%)	69013 (27.6%)	7282 (13.7%)	13922 (26.3%)	N/a	n (%)
Heart failure	81648 (26.5%)	27640 (47.6%)	54008 (21.6%)	6573 (12.4%)	10600 (20.0%)	N/a	n (%)
Diabetes mellitus	62392 (20.3%)	19120 (32.9%)	43272 (17.3%)	6581 (12.4%)	9407 (17.8%)	N/a	n (%)
Hypercholesterolemia	18549 (6.0%)	4736 (8.1%)	13813 (5.5%)	1738 (3.3%)	2580 (4.9%)	N/a	n (%)
Ischemic heart disease	92252 (30.0%)	24844 (42.8%)	67408 (27.0%)	9102 (17.2%)	12613 (23.8%)	N/a	n (%)
CKD <60 ml/min	65618 (21.3%)	21649 (37.3%)	43969 (17.6%)	7143 (13.5%)	9488 (17.9%)	N/a	n (%)
CKD <30 ml/min	10828 (3.5%)	4793 (8.2%)	6035 (2.4%)	1346 (2.5%)	1887 (3.6%)	N/a	n (%)
CRT-D	6071 (2.0%)	1820 (3.1%)	4251 (1.7%)	267 (0.5%)	663 (1.3%)	N/a	n (%)
ICD	38581 (12.5%)	9954 (17.1%)	28627 (11.5%)	2869 (5.4%)	3645 (6.9%)	N/a	n (%)
Laboratory							
NT-proBNP (LV)	406.0 [129.0–1302.0]	866.0 [271.0–2405.0]	311.0 [108.0–993.0]	317.0 [98.0–1247.0]	331.0 [100.0–1337.8]	47.9	ng/L
Troponin T (LV)	16.0 [8.0–44.0]	20.0 [10.0–50.0]	14.0 [7.0–41.0]	17.0 [7.0–58.0]	17.0 [8.0–53.0]	55.9	ng/L
eGFR (LV)	67.0 [59.0–87.0]	60.0 [48.0–81.0]	69.0 [60.0–88.0]	65.0 [60.0–88.0]	72.0 [60.0–90.0]	17.7	mL/min/1.73m ²
Creatinine	84.0 [71.0–102.0]	91.0 [74.0–119.0]	83.0 [70.0–99.0]	80.0 [68.0–96.0]	82.0 [69.0–99.0]	16.5	μmol/L
Sodium	140.0 [138.0–141.0]	139.0 [137.0–141.0]	140.0 [138.0–142.0]	140.0 [138.0–142.0]	140.0 [138.0–141.0]	21.4	mmol/L
Potassium	4.2 [4.0–4.5]	4.2 [4.0–4.6]	4.2 [4.0–4.5]	4.2 [3.9–4.4]	4.2 [3.9–4.5]	21.2	mmol/L
CRP	3.1 [1.5–12.0]	4.0 [1.9–13.9]	3.0 [1.4–11.3]	3.6 [1.7–14.0]	3.4 [1.5–14.2]	32.6	mg/L
Hemoglobin	8.5 [7.7–9.2]	8.3 [7.4–9.1]	8.5 [7.8–9.2]	8.5 [7.7–9.2]	8.4 [7.6–9.1]	18.2	mmol/L
Echocardiography							
LVEF	58.0 [50.0–67.0]	53.0 [42.0–64.0]	59.0 [52.0–68.0]	58.0 [51.0–67.0]	58.0 [51.0–67.0]	82.9	%
LVEF ≤40%	5996 (11.4%)	2404 (22.9%)	3592 (8.5%)	1402 (9.0%)	1402 (9.0%)	82.9	n (%)
Prior healthcare util							
Hosp last 14m	1.9 ± 1.8	2.4 ± 2.3	1.7 ± 1.4	1.4 ± 0.9	1.5 ± 1.2	73.3	n
Hosp contacts last 14m	7.7 ± 8.2	11.0 ± 10.1	6.9 ± 7.4	3.8 ± 6.1	6.1 ± 7.4	0.0	n
OPV / calls last 14m	3.7 ± 3.0	4.9 ± 4.0	3.4 ± 2.6	1.1 ± 0.4	2.7 ± 2.1	0.0	n
Medication							
ACEi/ARB/ARNI	111170 (36.1%)	30806 (53.0%)	80364 (32.2%)	7019 (13.2%)	15316 (28.9%)	N/a	n (%)
Beta-blocker	135137 (43.9%)	35238 (60.6%)	99899 (40.0%)	8101 (15.3%)	18622 (35.1%)	N/a	n (%)
MRA	49871 (16.2%)	17450 (30.0%)	32421 (13.0%)	2533 (4.8%)	6240 (11.8%)	N/a	n (%)
Loop diuretic	81203 (26.4%)	27085 (46.6%)	54118 (21.7%)	5500 (10.4%)	11077 (20.9%)	N/a	n (%)
SGLT2 inhibitor	10914 (3.5%)	4313 (7.4%)	6601 (2.6%)	486 (0.9%)	2110 (4.0%)	N/a	n (%)
Calcium channel blocker	92696 (30.1%)	24495 (42.1%)	68201 (27.3%)	6454 (12.2%)	13591 (25.6%)	N/a	n (%)
Thiazide	26155 (8.5%)	7571 (13.0%)	18584 (7.4%)	1768 (3.3%)	3728 (7.0%)	N/a	n (%)
Event rates							
1-year hospitalization	30586 (9.9%)						
2-year hospitalization	44419 (14.4%)						
1-year death	10783 (3.5%)						
2-year death	19809 (6.4%)						
1-year composite	38525 (12.5%)						
2-year composite	58701 (19.1%)						

1 Table 1: Baseline characteristics of all cardiology outpatient visits
2 included in the initial analytic cohort before model-specific

preprocessing steps. This cohort represents the full set of outpatient visits prior to applying follow-up requirements, exclusion of visits occurring <24 hours before an unplanned admission, and near-duplicate trigger deduplication used for machine-learning model training and validation. Event rates (1-year and 2-year hospitalization, death, and their composite) reflect unadjusted outcomes in the unfiltered baseline cohort. Subsequent model performance metrics and subgroup analyses are based on the filtered prediction cohort after applying these preprocessing steps. For binary comorbidity and medication flags, missingness is not applicable. Reported 2-year event rates may slightly underestimate true incidence due to incomplete follow-up of recent visits.

Model Performance

Discrimination

For prediction of the composite endpoint within 2 years, model performance was evaluated in the filtered, leakage-safe prediction cohort (N = 199,961) after application of all preprocessing and leakage-prevention steps described in the Methods. Across five patient-level GroupKFold validation folds, the mean AUROC was 0.77 ± 0.00 and the AUPRC was 0.42 ± 0.01 , given an overall 2-year event prevalence of 16.8%. Discrimination and threshold performance are illustrated in

Figure 1 A-C, displaying the ROC and PR curves together with the precision-recall-specificity profiles across thresholds.

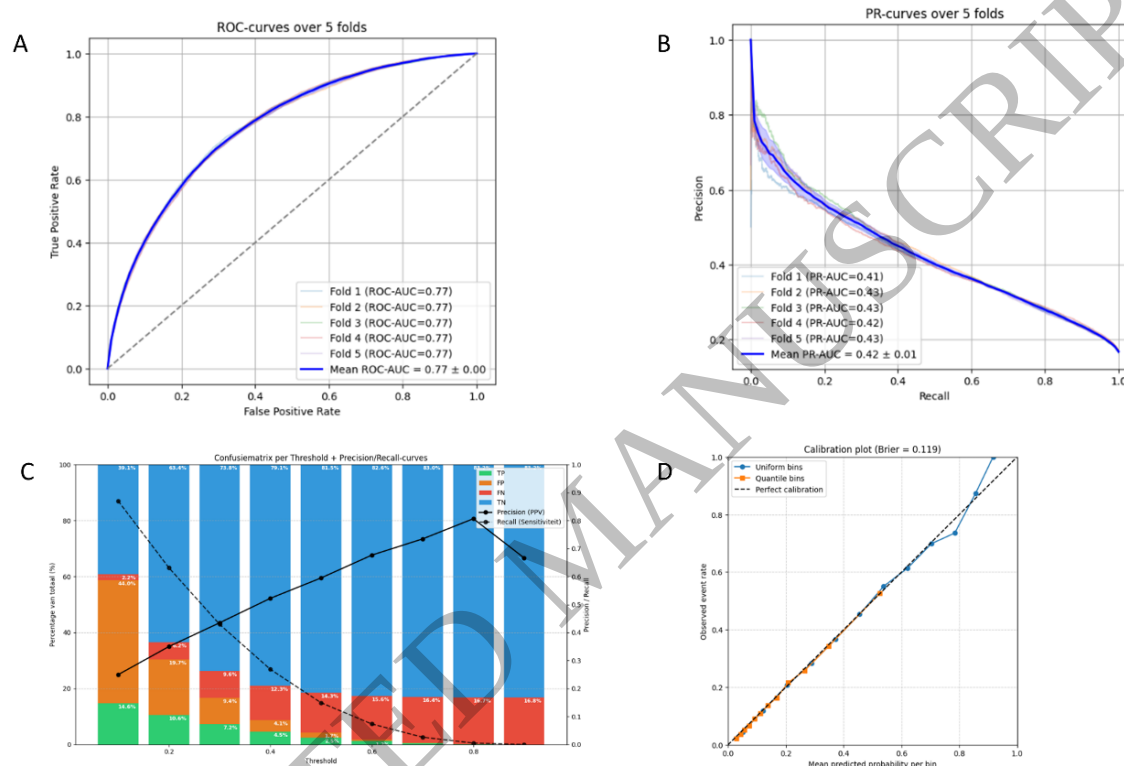


Figure 1: Discrimination and calibration performance of the CHARP model for predicting the 2-year composite endpoint of unplanned cardiac hospitalization or death. (A) Receiver operating characteristic (ROC) curves across five GroupKFold validation folds. (B) Precision-recall (PR) curves across the same folds. (C) Threshold analysis illustrating the trade-off between precision, recall, and class proportions across probability cut-offs. (D) Calibration plot showing agreement between predicted and observed event rates, with overall Brier score of 0.119.

Calibration

Calibration analysis showed close agreement between predicted and observed event rates across deciles of predicted risk. The calibration curve closely followed the ideal 45-degree line, indicating minimal systematic bias; additional isotonic or logistic recalibration was therefore not required during model development and internal validation. The Brier score was 0.119, reflecting good overall accuracy of the probabilistic predictions (Figure 1D).

Sensitivity analyses

Follow-up completeness sensitivity analysis: Requiring full 730-day follow-up for apparently event-free triggers reduced the prediction cohort from 199,961 to 193,200 outpatient visits, while model performance remained materially unchanged (AUROC 0.769 vs 0.768; AUPRC 0.419 vs 0.427; Brier score 0.118 vs 0.121), supporting

1 robustness of the primary findings to stricter follow-up completeness
2 requirements (Supplementary Appendix C).

3 Endpoint-specific sensitivity analyses: For 2-year unplanned cardiac
4 hospitalization alone, the model achieved an AUROC of 0.751 (SD
5 0.004) and an AUPRC of 0.312 (SD 0.014). For 2-year all-cause mortality
6 alone, AUROC was 0.873 (SD 0.004) and AUPRC was 0.335 (SD 0.013). A
7 shorter 1-year composite endpoint yielded similar discrimination
8 performance (AUROC 0.766, SD 0.002; AUPRC 0.309, SD 0.007)
9 compared with the primary 2-year composite endpoint (AUROC 0.769,
10 SD 0.003; AUPRC 0.418, SD 0.006) (Supplementary Appendix D).

11 Benchmark comparison with logistic regression: Compared with
12 regularized logistic regression using the same feature set and validation
13 strategy, XGBoost achieved superior performance (AUROC 0.77 vs 0.74;
14 AUPRC 0.42 vs 0.39; Brier score 0.118 vs 0.188), supporting the added

1 value of nonlinear modeling and feature interactions in structured EHR
2 data.

3 Feature Importance and Explainability

4 Feature importance was evaluated across all validation folds using

5 XGBoost's intrinsic metrics: Gain, representing the average

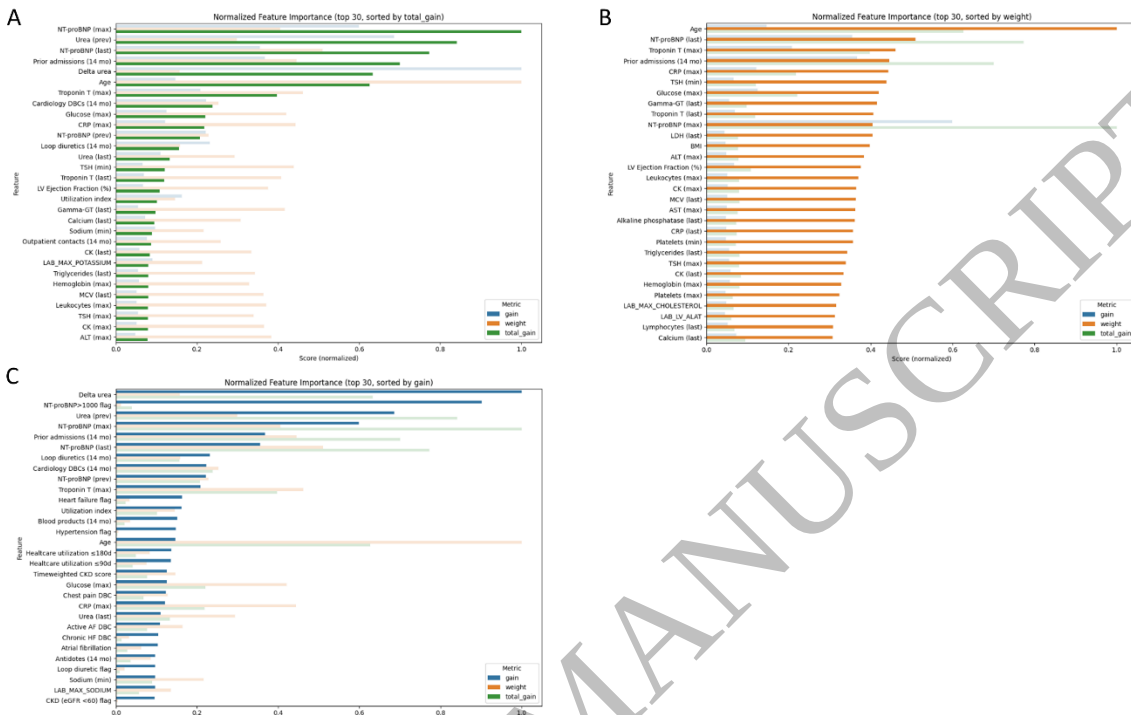
6 improvement in model performance (information gain) brought by a

7 feature when used for splitting; Total gain, the sum of all gains for a

8 given feature across all trees; Weight, the number of times a feature

9 was used for splitting within the model. The top 30 features ranked by

1 these metrics are shown in Figure 2



2
3 Figure 2: Top 30 model features ranked by three XGBoost importance
4 metrics: (A) Total Gain (cumulative improvement in model
5 performance), (B) Weight (frequency of use in tree splits), and (C)
6 Gain (average improvement per split).

7
8 Across all three measures, variables related to NT-proBNP and renal
9 function consistently ranked highest. Maximum NT-proBNP, NT-
10 proBNP>1000 flag, and recent NT-proBNP values were the most

1 influential predictors, followed by serum urea levels (maximum, before-
2 last, and delta values) and the presence of chronic kidney disease.

3 Indicators of healthcare utilization, including prior admissions,
4 outpatient contacts, and encounter counts in the preceding months,
5 also contributed substantially to model performance. Additional
6 features with notable importance included troponin T, C-reactive
7 protein, serum glucose, age, and flags for heart failure, hypertension,
8 and reduced LVEF $\leq 40\%$. Local model explainability was assessed using
9 Shapley Additive Explanations (SHAP) to quantify, for each individual
10 prediction, the direction and magnitude of each feature's contribution
11 to the predicted risk. For every outpatient visit ("trigger"), SHAP values
12 were computed to identify which variables most strongly increased or
13 decreased the probability of cardiac hospitalization or death. During
14 model deployment, the five highest-impact features (based on absolute
15 SHAP values) are displayed within the electronic medical record
16 alongside the predicted probability, enabling clinicians to understand

- 1 why a specific patient is flagged as higher or lower risk. Figure 3
- 2 illustrates an example of an outpatient visit with a predicted 2-year risk
- 3 of 72%, including the corresponding SHAP-based feature contributions.

Outpatient planning – 07 April 2025

K...	Tijd	CHARP score	Patient	Gewicht	Polsf...	Last BP	Laatste e...	Type spr.	Type	Status
●	09:00	N.v.t.*	Mendel, C 62 jr. / V	81 kg	88	(I) 125/92		Card Nieuwe Patient Vumc Algemeen	CARD NIEUWE PATIENT VUMC	Getekend Ingecheckt: 08:28
●	09:20	32%*	Sapher-Stro, JJ 74 jr. / M	99 kg	90	126/84 58		Card Cp Vumc	CARD CP VUMC	Getekend Ingecheckt: 09:20
●	09:40	27%*	Grehmal, CA 81 jr. / M	79 kg	58	135/70 56		Card Cp Vumc	CARD CP VUMC	Getekend Ingecheckt: 09:35
●	10:00	10%*	Terwinde, O 56 jr. / M	95 kg	52	133/76 81		Card Cp Vumc	CARD CP VUMC	Getekend Ingecheckt: 09:55
○	10:20	25%*	Restilinga, P 87 jr. / M	84 kg	60	138/75 67		Card Cp Vumc	CARD CP VUMC	Niet gezien No-show
●	10:40	9%*	Pot, V 79 jr. / V	79 kg	64	(I) 178/95 75		Card Cp Vumc	CARD CP VUMC	Getekend Ingecheckt: 10:31
●	11:00	55%*	Querto, SL 74 jr. / M	102 kg	63	110/71 53		Card Cp Vumc	CARD CP VUMC	Getekend Ingecheckt: 10:47
●	11:20	53%*	Freddeling-Seh, J 67 jr. / V	83 kg	84	145/86 88		Card Cp Vumc	CARD CP VUMC	Getekend Ingecheckt: 10:08
○	12:00	54%*	Dissel-Sert, AA	88 kg	82	(I) 122/82 80		Card Telefonisch Consult Vumc	CARD TELEFONISCH CONSULT VUMC	Getekend Ingecheckt: 07:59
○	12:15	25%*						isch Consult Vumc	CARD TELEFONISCH CONSULT VUMC	Getekend Ingecheckt: 07:59
○	12:30	12%*						isch Consult Vumc	CARD TELEFONISCH CONSULT VUMC	Getekend Ingecheckt: 07:59
○	12:45	46%*						isch Consult Vumc	CARD TELEFONISCH CONSULT VUMC	Getekend Ingecheckt: 07:59
○	13:00	78%*						inc	CARD TEL VUMC	Getekend Ingecheckt: 07:59
○	13:15	72%*						isch Consult Vumc complex	CARD TELEFONISCH CONSULT VUMC	Getekend Ingecheckt: 07:59

UMCA PA CHARP 2.0

[Details historische scores en model verkennen](#)

Score niet voor medisch gebruik. Model is in status validatie.

72%

Feature 1: NT-proBNP (31%)
Feature 2: Prior Admissions (14mo) (22%)
Feature 3: Age (13%)
Feature 4: eGFR (12%)
Feature 5: Urea (8%)

Hoog risico: 60 - 100
Gemiddeld risico: 20 - 60
Laag: poli over 2 jr: 0 - 20

- 4
- 5 Figure 3: Example of a cardiology specialist's outpatient agenda
- 6 displaying a single highlighted visit with a predicted 2-year risk of
- 7 unplanned cardiac hospitalization or death of 72%, including the
- 8 corresponding SHAP-based feature contributions for that visit.
- 9 Feature contributions represent local explanations relative to the
- 10 model's expected value. Patient identifiers, including names and
- 11 ages, have been modified for privacy purposes. At the time of
- 12 analysis, the model was deployed in silent-running mode, with
- 13 predictions and explanations visible to researchers only and not used
- 14 for clinical decision-making. Image adapted for research purposes.
- 15 Copyright © Amsterdam UMC. Epic® is a registered trademark of Epic
- 16 Systems Corporation.

Subgroup Analysis

Model performance was similar across all major clinical subgroups (Table 2). Discrimination and calibration metrics showed limited variation between subgroups. AUPRC was higher in subgroups with higher event rates, including chronic kidney disease and heart failure. Overall performance metrics remained comparable across the evaluated cardiovascular comorbidities.

Subgroup	N visits	Events	Event rate	AUROC	AUPRC	Brier
CKD (eGFR<30)	6430	2674	0.42	0.74	0.64	0.20
Heart failure	47249	14752	0.31	0.72	0.53	0.19
CKD (eGFR<60)	39522	12074	0.31	0.73	0.54	0.18
Diabetes	38013	10774	0.28	0.73	0.52	0.17
Age ≥70y	69003	17297	0.25	0.72	0.48	0.16
Ischemic heart disease	56674	14087	0.25	0.73	0.48	0.16
Hypertension	54953	12837	0.23	0.74	0.47	0.15
All patients	199961	33627	0.17	0.77	0.42	0.12

Table 2: Out-of-fold model performance across major clinical subgroups, showing event rates and discrimination metrics (AUROC and AUPRC) together with overall probabilistic accuracy expressed as the Brier score. Visit counts reflect the filtered prediction cohort used for out-of-fold validation and are therefore lower than the unfiltered baseline cohort shown in Table 1. Differences in Brier score across subgroups largely reflect differences in event prevalence and should not be interpreted as subgroup-specific

1 miscalibration. Overall model calibration was assessed in the full
2 cohort.

3 Generalizability and Cross-site Performance

4 After preprocessing, model performance was evaluated across four
5 validation settings. In the site-exclusive experiments, training on VUMC
6 data and testing on AMC data resulted in an AUROC of 0.737 and an
7 AUPRC of 0.360, while training on AMC data and testing on VUMC data
8 yielded an AUROC of 0.725 and an AUPRC of 0.420. In the patient-
9 exclusive internal evaluations, where 15% of patients per site were
10 reserved as a holdout set, training on the combined dataset excluding
11 15% of VUMC patients and testing on the VUMC holdout cohort
12 resulted in an AUROC of 0.748 and an AUPRC of 0.456. Similarly,
13 training on the combined dataset excluding 15% of AMC patients and
14 testing on the AMC holdout cohort achieved an AUROC of 0.786 and an
15 AUPRC of 0.417.

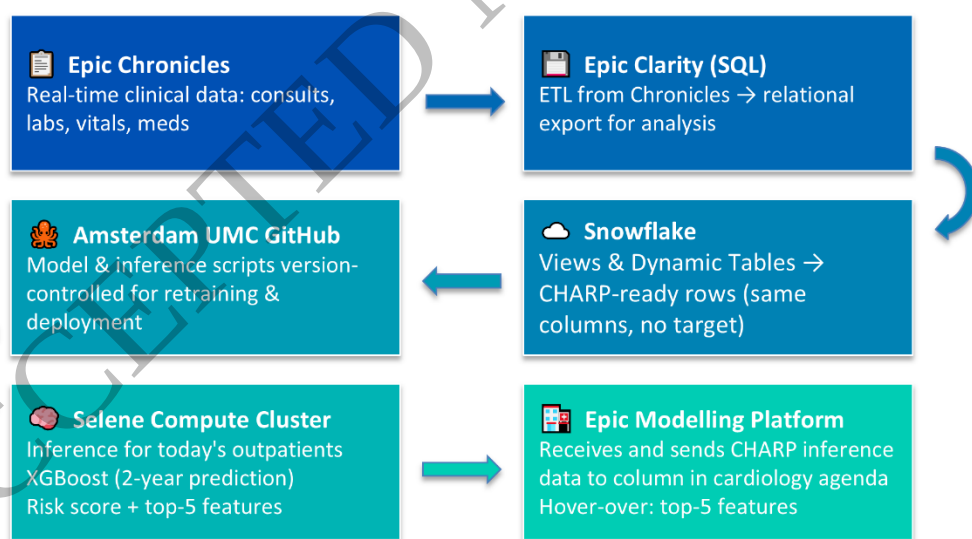
16 Model Deployment

17 The CHARP model was fully integrated within the hospital's Epic
18 environment and data platform. Clinical data originate in Epic

Chronicles, Epic's non-relational transactional database that underpins the live EHR. For analytical and research use, data are replicated into Epic Clarity, a relational database optimized for reporting, and subsequently exposed through curated Snowflake views representing the outpatient cardiology cohort and derived features used by the model. Each morning at 06:00, all patients scheduled for the outpatient cardiology clinic that day are automatically retrieved from Epic. Their data are transformed using the same feature-engineering logic as in model training and sent to the CHARP inference cluster, a containerized service version-controlled in Git. The resulting 2-year composite risk predictions are returned through Epic's Cognitive Computing / Nebula predictive-modeling framework and written back into the operational Epic environment, as illustrated in Figure 4. Within Epic Hyperspace, the risk score appears as an additional column in the outpatient clinic agenda. A hover-over tooltip displays the five most influential features for each prediction, providing transparent and interpretable decision

1 support. The system runs fully automatically, without manual input,
 2 and scores all scheduled patients once daily.

3 The present study reports retrospective model development and
 4 validation results only. Although the model has been technically
 5 deployed in a prospective silent-running environment, outputs
 6 generated during this phase were not used to guide clinical decision-
 7 making and were not included in the current analyses. Prospective
 8 silent-running evaluation results will be reported separately.



Manual model retraining every 6 months on Selene using new Snowflake data (via GitHub)

9

10 Figure 4: System architecture of the CHARP pipeline, illustrating data
 11 flow from real-time clinical sources in Epic to inference and

deployment across Snowflake, Selene Compute Cluster, and the Epic Modeling Platform, with planned biannual model retraining managed via GitHub.

Risk of Bias and Applicability

Using the updated PROBAST+AI framework (20), the CHARP model

demonstrated a low overall risk of bias and high applicability to

outpatient cardiology. Participants: The cohort consisted of all

cardiology outpatients across two academic hospitals within one

healthcare system. This ensures internal consistency but may limit

generalizability, consistent with a moderate applicability concern due

to the single-system setting. Predictors: All predictors were routinely

collected EHR variables available at, or before, the outpatient visit.

Extensive leakage-prevention measures, removal of raw timestamps,

exclusion of <24h pre-admission data, patient-level splits, and

deduplication of near-duplicate visits, yielded a low risk of predictor-

related bias, including AI-specific leakage domains introduced in

PROBAST+AI (temporal, sampling, and pipeline leakage). Outcome:

Outcomes were defined using structured internal hospital codes and

1 municipal mortality linkage. A residual limitation is incomplete capture
2 of unplanned cardiac hospitalizations in external hospitals, leading to
3 possible under-ascertainment of true events. PROBAST+AI classifies this
4 as a potential outcome-misclassification risk. Analysis: The use of
5 patient-level GroupKFold, strict feature-timestamp masking, gradient
6 boosting with robust missing-handling, and calibration assessment
7 aligns with PROBAST+AI's recommendations for AI-based prediction
8 models. Overall analysis-related bias was low, with adequate sample
9 size, transparent hyperparameter optimization, and reproducible
10 pipelines. Applicability: Applicability to similar EHR-rich cardiology
11 settings is high; external validation in independent systems will be
12 required to mitigate the moderate applicability concern arising from
13 single-system derivation.

Discussion

In this large-scale study of over 300,000 outpatient cardiology visits across two academic centers, we developed and validated a machine learning (ML) model, CHARP, to predict unplanned cardiac hospitalization or death within two years after an outpatient visit. The model achieved robust discrimination (AUROC 0.77; AUPRC 0.42) and good calibration (Brier 0.12), while maintaining consistent performance across clinically relevant subgroups and across the two historically distinct Amsterdam UMC hospital sites, supporting transportability across different clinical environments within the same healthcare system. These findings show that routinely collected electronic health record (EHR) data can be used to estimate individualized risk for adverse cardiac outcomes and may support identification of patients at very low risk in whom routine follow-up intensity could be reduced.

Comparison with prior studies

Previous machine learning (ML) approaches to hospitalization or readmission risk prediction have typically shown modest performance

(AUROC 0.65–0.75) and limited transportability beyond their derivation cohorts.(4–14) However, many existing clinical models focus specifically on short-term readmission risk, typically predicting 30-day rehospitalization following hospital discharge, rather than longer-term risk of unplanned hospitalization and death from the outpatient setting. As such, direct comparison with these models is inherently limited due to differences in prediction target, clinical context, and time horizon. CHARP’s discriminative ability nevertheless compares favorably with these reports and with contemporary readmission models used in clinical operations, such as the Epic Readmission Risk Model and the LACE index, which generally perform in the 0.65–0.70 range and are designed to estimate short-term readmission risk rather than longer-term hospitalization risk.(18) In contrast, CHARP predicts the risk of unplanned cardiac hospitalization or death over a two-year time horizon, reflecting a different clinical use case focused on longitudinal outpatient risk stratification. The gain in accuracy likely reflects the comprehensive feature engineering strategy, the use of gradient

boosting (XGBoost) with extensive hyperparameter optimization, and rigorous prevention of data leakage at both feature and patient levels. Importantly, model calibration, often neglected in high-dimensional ML studies, was good in the current internal validation setting without the need for immediate post-hoc recalibration, supporting clinical interpretability and safe use for probabilistic risk estimation.

Key predictors and pathophysiological Insights

The strongest predictors of cardiac hospitalization and death were NT-proBNP, markers of renal function (eGFR, urea, creatinine), and prior healthcare utilization. This aligns with known cardio-renal interactions, where elevated natriuretic peptides and reduced renal function jointly indicate hemodynamic stress, volume overload, and systemic vulnerability. Other influential variables included troponin T, C-reactive protein, age, and prior heart failure, all consistent with established prognostic determinants in chronic cardiovascular disease. Notably, sex did not emerge as a major independent predictor. Global predictor importance was primarily derived from native XGBoost importance

metrics, while SHAP values were additionally used to support visit-level explainability within the EHR interface. Importantly, both approaches identified broadly similar dominant predictors. The model's explainability displayed directly in the electronic medical record, allows clinicians to understand these relationships at the individual-patient level, which may foster trust and facilitate adoption.(21)

Clinical and Operational Implications

The CHARP program aims to support risk-informed outpatient care by identifying patients at low risk for hospital admission or death in whom follow-up intervals may be safely extended, thereby reducing unnecessary routine visits while maintaining existing care pathways and clinical oversight. Predictions reflect observed outcomes under current care pathways rather than underlying untreated risk. Therefore, in individual patients, treating physicians may decide that intensified follow-up itself contributes to clinical stability and should therefore be maintained. At the same time, one of the central hypotheses underlying the CHARP program is that a substantial proportion of

1 routine outpatient follow-up visits in current practice may not
2 materially alter short- to medium-term outcomes. Prospective
3 evaluation will therefore be necessary to determine whether model-
4 informed follow-up de-intensification can be implemented safely
5 without unintended effects on patient outcomes.

6 When validated prospectively, such risk-informed follow-up strategies
7 could alleviate pressure on outpatient capacity without compromising
8 safety. Given that outpatient cardiology visits represent one of the
9 highest-volume services in European tertiary care, even modest
10 efficiency gains could translate into substantial system-level impact.

11 Although the present study reports retrospective model development
12 and internal validation only, the model has already been technically
13 deployed in a prospective silent-running environment fully integrated
14 within the Epic EHR. In this setting, predictions are generated
15 automatically for all scheduled outpatients but are not yet visible and
16 for clinicians and not yet used to support clinical decision-making. This

1 infrastructure enables future prospective evaluation of real-world
2 model behavior, calibration drift, and clinician interaction prior to
3 implementation of live decision support. Importantly, real-world EHR
4 integration and deployment infrastructure remain insufficiently
5 described in the current prediction-model literature, despite being
6 critical prerequisites for prospective clinical evaluation and
7 implementation. The overall architecture therefore reflects a learning
8 health-system framework in which model performance, data quality,
9 and outcomes can be continuously monitored and evaluated.

10 Strengths and Innovations

11 This study combines several methodological and translational
12 strengths, including a harmonized longitudinal EHR infrastructure
13 across two historically distinct hospital sites within one academic
14 healthcare system using standardized SNOMED-CT and LOINC
15 terminology; a feature-engineering framework capturing healthcare
16 utilization, biomarker trajectories, and temporal clinical patterns;
17 extensive safeguards against temporal and patient-level leakage,

including strict patient-level separation and near-duplicate visit handling during validation; full integration into the operational EHR infrastructure for automated daily visit-level risk estimation; trigger-level explainability within the EHR using SHAP-based feature attribution; and an implementation-oriented design focused on identifying patients at very low risk to support safe de-intensification of routine outpatient follow-up. Together, these features position CHARP as a scalable and implementation-ready framework for prospective evaluation of AI-supported outpatient care optimization.

Limitations

Several limitations warrant consideration. First, although the dataset was comprehensively harmonized, it originates from a single evolving healthcare system encompassing two historically distinct tertiary hospital sites. Therefore, true external validation in independent healthcare systems remains necessary to assess broader transportability and clinical applicability. However, for complex, high-dimensional EHR-based prediction models such as CHARP, strict cross-

1 institution generalizability may not always be the primary objective.

2 Local use and retraining or recalibration in external hospitals may yield
3 superior discrimination by capturing institution-specific clinical
4 practices, case mix, and data-generation processes.(16) This raises an
5 important conceptual trade-off between global generalizability and
6 locally optimized predictive performance, which should be explicitly
7 addressed in future deployment studies.

8 Second, although the endpoint already included a broad range of acute
9 cardiac and critical care departments, certain clinically related non-
10 cardiac admissions may still have been missed, particularly when
11 managed without cardiology involvement or when occurring at
12 hospitals outside Amsterdam UMC

13 Third, the model relies exclusively on structured EHR variables.

14 Unstructured clinical notes, raw imaging data, and genomic information
15 were not incorporated. While this design choice enhances transparency
16 and facilitates deployment, it likely constrains the upper bound of

1 predictive performance. Future multimodal extensions integrating
2 natural language features and imaging-derived biomarkers may further
3 improve discrimination and clinical relevance.

4 Fourth, although XGBoost natively accommodates missing values and
5 may leverage predictive information contained in missingness patterns,
6 this does not resolve the reasons why values are missing, particularly
7 when data are missing not at random. In EHR-derived datasets,
8 missingness itself may encode clinically meaningful or operational
9 information, such as selective laboratory ordering or workflow-related
10 data availability. These patterns may also vary across institutions and
11 could affect model transportability if not explicitly evaluated or
12 recalibrated. Furthermore, although XGBoost provides strong predictive
13 flexibility, more interpretable models such as logistic regression may be
14 preferable when mechanistic interpretation or targeted interventions
15 are the primary objective. Fifth, the current implementation operates in
16 silent-running mode. Consequently, the impact of CHARP on clinician

behavior, workload, decision-making, and patient outcomes has not yet been evaluated. Prospective studies, including randomized or stepped-wedge designs, will be necessary to assess real-world clinical effectiveness, safety, and unintended consequences prior to routine clinical use.

Sixth, because the primary endpoint was designed to reflect overall cardiology outpatient vulnerability and healthcare demand rather than a single disease-specific cardiovascular process, CHARP should primarily be interpreted as a pragmatic outpatient risk-stratification tool for follow-up optimization rather than a disease-specific prognostic model.

Seventh, although strict patient-level separation and 90-day near-duplicate visit deduplication were applied during validation, residual clustering at provider-, clinic-, or hospital-level may still persist.

Finally, deployment of CHARP may influence clinical behaviour, potentially altering predictor distributions and event incidence over time. This may result in calibration drift despite stable discrimination,

1 highlighting the need for prospective monitoring and periodic
2 recalibration during implementation.

3 Future Directions

4 The next phases of the CHARP program will follow a staged translational
5 roadmap. First, prospective silent-running validation will be conducted
6 to assess real-time model performance, calibration drift, and temporal
7 robustness in routine clinical care. Second, CHARP will be implemented
8 as a decision support tool within the clinical workflow, explicitly designed
9 to support, rather than replace, clinical judgment. This phase will focus
10 on safe integration with full clinician override, evaluation of usability and
11 interpretability, and assessment of potential behavioral effects on
12 clinical decision-making. Third, a randomized clinical trial will evaluate
13 the clinical impact of CHARP-guided care, including effects on outpatient
14 visit frequency, unplanned hospitalizations and mortality. In parallel,
15 existing explainability dashboards and user-interface components,
16 already implemented within the current CHARP infrastructure, will be
17 iteratively optimized in close collaboration with clinicians to further

1 enhance transparency, usability, and trust. Equally important, future
2 work will focus on establishing robust model governance and lifecycle
3 management, including version control, performance monitoring,
4 recalibration strategies, and clearly defined ownership of long-term
5 maintenance, which are essential for safe and sustainable deployment in
6 routine care. Lastly, model improvement will be achieved through
7 systematic data enrichment through accumulation of additional
8 outpatient data and inclusion of newly available variables. Periodic
9 retraining (e.g. every year) will account for temporal changes in case-mix
10 and clinical practice, while incorporation of longitudinal biomarker
11 trajectories, healthcare utilization dynamics, and structured imaging-
12 derived features may improve discrimination and calibration. All updates
13 will be evaluated using strict temporal holdout testing within the same
14 leakage-prevention framework to ensure real-world validity. Beyond the
15 specific CHARP implementation, similar AI-supported risk stratification
16 approaches may also be applicable to other clinical settings, such as post-
17 emergency department follow-up pathways and additional outpatient

1 specialties where routine follow-up schedules may contribute to
2 unnecessary healthcare utilization.

3 Conclusion

4 The CHARP model provides a scalable, fully EHR-integrated approach
5 for visit-level risk estimation in outpatient cardiology. The model
6 demonstrated good discrimination, enabling identification of patients
7 at very low risk, and allows prospective evaluation of risk-informed
8 follow-up strategies aimed at safely reducing routine follow-up
9 intensity while maintaining quality and safety of care.

Data availability

The data underlying this article cannot be shared publicly due to patient privacy regulations, institutional data governance policies, and the use of protected electronic medical record data. De-identified data may be shared on reasonable request to the corresponding author, subject to approval by Amsterdam UMC and applicable legal and institutional governance procedures.

Ethics statement

The study protocol was reviewed by the Medical Ethics Committee of Amsterdam UMC and was classified as not subject to the Medical Research Involving Human Subjects Act (WMO). The requirement for informed consent was therefore waived. All data were pseudonymized

1 and processed in compliance with the General Data Protection

2 Regulation (GDPR).

3 Funding

4 No external commercial funding was received for the development or
5 evaluation of the CHARP model.

6 Author contributions

7 J.L. Selder conceived the study, led the project, and drafted the
8 manuscript. All authors contributed to study design, interpretation of
9 results, and/or critical revision of the manuscript. All authors approved
10 the final version of the manuscript.

11 Conflict of interest

12 The authors declare no conflicts of interest related to this work.

13 Consent

14 The need for informed consent was waived by the institutional Medical
15 Ethics Review Committee, as this retrospective study used

pseudonymized routinely collected clinical data and did not fall under the scope of the Dutch Medical Research Involving Human Subjects Act (WMO).

References

1. World Health Organization. Health workforce: Data and statistics. WHO Regional Office for Europe. Published 2020. Accessed November 2, 2020. <https://www.euro.who.int/en/health-topics/Health-systems/health-workforce/data-and-statistics>
2. Bates S, Vaz C, Ward M. Trendwatch Chartbook 2018. American Hospital Association; 2018.
3. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94–98. doi:10.7861/fhj.2019-0015
4. Angraal S, Mortazavi BJ, Gupta A, et al. Machine learning prediction of mortality and hospitalization in heart failure with

1 preserved ejection fraction. JACC Heart Fail. 2020;8(1):12–21.

2 doi:10.1016/j.jchf.2019.09.013

3 5. Bertsimas D, Pauphilet J, Stevens J, Tandon M. Predicting
4 inpatient flow at a major hospital using interpretable analytics.

5 Manufacturing & Service Operations Management.

6 2021;23(4):891–909. doi:10.1287/msom.2020.0921

7 6. Safavi KC, Khaniyev T, Copenhaver M, et al. Development and
8 validation of a machine learning model to aid discharge processes
9 for inpatient surgical care. JAMA Network Open.

10 2019;2(12):e1917221. doi:10.1001/jamanetworkopen.2019.17221

11 7. Desautels T, Das R, Calvert J, et al. Prediction of early unplanned
12 intensive care unit readmission in a UK tertiary care hospital: a
13 cross-sectional machine learning approach. BMJ Open.

14 2017;7(9):e017199. doi:10.1136/bmjopen-2017-017199

15 8. Golas SB, Shibahara T, Agboola S, et al. A machine learning model
16 to predict the risk of 30-day readmissions in patients with heart

1 failure: a retrospective analysis of electronic medical records data.

2 BMC Medical Informatics and Decision Making. 2018;18:44.

3 doi:10.1186/s12911-018-0620-z

4 9. Hilton CB, Milinovich A, Felix C, et al. Personalized predictions of
5 patient outcomes during and after hospitalization using artificial
6 intelligence. npj Digital Medicine. 2020;3:51. doi:10.1038/s41746-
7 020-0249-z

8 10. Hong WS, Haimovich AD, Taylor RA. Predicting hospital
9 admission at emergency department triage using machine
10 learning. PLoS One. 2018;13(7):e0201016.
11 doi:10.1371/journal.pone.0201016

12 11. Morawski K, Dvorkis Y, Monsen CB. Predicting
13 hospitalizations from electronic health record data. American
14 Journal of Managed Care. 2020;26(1):e17–e23.

15 12. O'Donovan F. Machine learning generated risk model to
16 predict unplanned hospital admission in heart failure. Circulation.

2017;136(Suppl 1):A16855.

13. Rahimian F, Salimi-Khorshidi G, Payberah AH, et al.

Predicting the risk of emergency admission with machine learning: development and validation using linked electronic health records. PLoS Medicine. 2018;15(11):e1002695.

14. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. 2018;1:18. doi:10.1038/s41746-018-0029-1

15. Fine B, Dharmarajan K, Wang Y, et al. Machine learning for risk prediction after emergency department visits or hospitalization for heart failure. PLOS Digital Health. 2024;3(2):e0000431. doi:10.1371/journal.pdig.0000431

16. de Hond AAH, van der Veen A, Bosman RJ, et al. External validation of the Pacmed Critical ICU readmission and mortality prediction model. Critical Care Medicine. 2023;51(1):e1–e11. doi:10.1097/CCM.0000000000005724

- 1 17. Al Alawi AM, Al Rawahi H, Al Harthi H, et al. Development
2 and evaluation of machine learning models to predict 30-day
3 hospital readmission. *Healthcare (Basel)*. 2025;13(1):112.
4 doi:10.3390/healthcare13010112
- 5 18. Hwang E, Lee S, Kim YJ, et al. External validation of Epic's risk
6 of unplanned readmission model. *PLoS One*.
7 2021;16(3):e0248980. doi:10.1371/journal.pone.0248980
- 8 19. Sabouri S, Heidari-Bateni G, Khosravi A, et al. Machine
9 learning-based readmission and mortality prediction in heart
10 failure patients. *Scientific Reports*. 2023;13:11421.
11 doi:10.1038/s41598-023-38292-9
- 12 20. Moons KGM, Wolff RF, Riley RD, et al. PROBAST+AI: A
13 framework for assessing risk of bias in prediction models using
14 artificial intelligence. *BMJ*. 2025;388:e079011. doi:10.1136/bmj-
15 2024-079011
- 16 21. Lundberg SM, Lee SI. A unified approach to interpreting

model predictions. *Advances in Neural Information Processing Systems*. 2017;30:4765–4774.

Appendix A. Workflow of the CHARP Prediction Pipeline

Overview of data flow, preprocessing, model development, evaluation, and deployment.

1. Data extraction and harmonization

Clinical data were extracted from curated Snowflake views representing the outpatient cardiology population. Extracted domains included demographics, vital signs, laboratory results, medications, procedures, encounters, and imaging metadata (modality and timestamp only; no pixel data). Variables were harmonized across Epic systems using standardized LOINC and SNOMED CT mappings.

2. Preprocessing and leakage prevention

To prevent information leakage, several safeguards were applied prior to model training:

- Masking and exclusion of data occurring ≤ 24 hours before an unplanned admission.
- Deduplication of near-duplicate outpatient triggers occurring within 90 days.
- Removal of all raw timestamp variables (*_DATE) after feature derivation.

3. Feature engineering

Derived features captured both clinical status and longitudinal care patterns:

- Healthcare utilization: exponential-decay utilization index (UTIL_INDEX), encounter counts over 30/90/180 days (UTIL_COUNT), Diagnosis Treatment Combination (DBC) counts, and medication counts.
- Laboratory and vital signs: last value (LV), before-last value (BLV), delta (LV-BLV), and selected minimum/maximum values.
- Renal function: chronic kidney disease (CKD) stage and a time-weighted CKD score (CKD_SCORE_TIMEWEIGHTED).

- Cardiac function: left ventricular ejection fraction (LVEF) extracted using a custom script from echocardiography reports and vendor databases.
- Procedures and imaging: counts of echocardiography, coronary angiography, and other structured procedures.

4. Model development

Risk prediction was performed using gradient boosting (XGBoost) with native handling of missing values. Hyperparameters were optimized using Optuna. Model development and internal validation employed patient-level GroupKFold cross-validation to ensure complete separation of individuals across folds.

5. Model evaluation and explainability

Performance was assessed using AUROC, AUPRC, calibration plots, and Brier score.

Explainability was implemented at two levels:

- Trigger-level: SHAP values identifying the five most influential features for each individual prediction.
- Model-level: XGBoost feature importance metrics (gain, weight, and total gain).

Deployment (silent run)

The trained model was deployed on a high-performance computing cluster positioned between Snowflake and the Epic EHR. The pipeline runs daily and generates 2-year risk estimates for all scheduled outpatient cardiology visits. During the silent-running phase, predictions and SHAP explanations are displayed in the EHR for

monitoring and validation purposes only and are not used to guide clinical decision-making.

Appendix B. Definitions of Key Clinical Variables in Snowflake

Hypertension

Hypertension (HT_FLAG) was defined using three independent electronic health record signals:

1. Diagnosis codes: hypertension-related Diagnosis Treatment

Combinations (DBC: 0320.902, 0313.311, 0316.4003, 8416.0107,

8416.0106, 0318.901), excluding pulmonary and ocular
hypertension.

2. Blood pressure: mean arterial pressure (MAP;
VIT_LV_BLOOD_PRESSURE) ≥ 110 mmHg, corresponding
approximately to systolic/diastolic values of 150/100 mmHg.

This threshold was selected to reduce misclassification due to
transient hospital-related blood pressure elevation.

3. Medication: antihypertensive therapy weighted by class:
thiazides, calcium channel blockers, and $\alpha 1$ -blockers (1 point);

1 RAAS inhibitors and β -blockers (0.5 points); loop diuretics (0
2 points). A cumulative score ≥ 2 classified treated hypertension.
3 Patients were classified as hypertensive if any of these criteria were
4 met.

5 Heart Failure

6 Heart failure (HF_FLAG) was defined using a composite of diagnostic,
7 functional, biomarker, device, and medication-based signals:

- 8 1. Diagnosis codes: heart failure-related DBCs (e.g., 0320.301,
9 0320.302, 0320.303, 0320.305, 0320.306, 8416.0105).
- 10 2. Cardiac function: LVEF $< 40\%$ on echocardiography or cardiac MRI.
- 11 3. Biomarker: NT-proBNP > 1000 ng/L.
- 12 4. Device therapy: presence of an ICD or CRT-D (implantation,
13 replacement, or follow-up).
- 14 5. Medication: guideline-directed heart failure therapy (RAAS
15 inhibitor/ARNI, β -blocker, MRA, loop diuretic, SGLT2 inhibitor).

1 Patients using ≥ 4 drug classes, or an ICD plus ≥ 3 classes, were
2 classified as having heart failure.

3 Patients meeting any criterion were classified as having heart failure.

4 Diabetes Mellitus

5 Diabetes mellitus (DM_FLAG) was defined using three independent EHR
6 signals:

- 7 1. Diagnosis codes: diabetes-related DBCs (e.g., 0335.222, 0318.902,
8 0313.223, 0316.7113, 0313.222, 0316.7114, 0313.221, 0305.2065,
9 0303.432), excluding gestational diabetes and diabetes insipidus.
- 10 2. Laboratory criteria: HbA1c ≥ 42 mmol/mol or plasma glucose ≥ 18
11 mmol/L, excluding non-blood specimens and point-of-care
12 measurements.
- 13 3. Medication: use of any glucose-lowering therapy, including
14 insulin, metformin, sulfonylureas, DPP-4 inhibitors, GLP-1
15 receptor agonists, SGLT2 inhibitors, or thiazolidinediones.

16 Patients were classified as diabetic if any criterion was present.

17 Hypercholesterolemia

Hypercholesterolemia (HC_FLAG) was defined by:

- Diagnosis codes: DBCs explicitly indicating hypercholesterolemia or dyslipidemia.
- Laboratory criteria: LDL cholesterol ≥ 5.0 mmol/L or total cholesterol ≥ 8.0 mmol/L, thresholds chosen to identify severe or familial hypercholesterolemia.

Lipid-lowering therapy was not used as a defining criterion, given its widespread use for secondary prevention in cardiovascular patients.

Ischemic Heart Disease

Ischemic heart disease (IHD_FLAG) was defined using a composite of diagnosis-based, procedure-based, and biomarker-based electronic health record signals, evaluated relative to each outpatient visit.

1. Diagnosis codes (DBC)

Patients were flagged as having ischemic heart disease if a prior Diagnosis Treatment Combination (DBC) indicative of ischemic heart disease was present before the outpatient visit. Included DBCs covered stable angina pectoris, unstable angina, ST-

elevation myocardial infarction (STEMI), non–ST-elevation myocardial infarction (NSTEMI), and other ischemic heart disease–related diagnoses. For each patient and outpatient visit, the most recent qualifying DBC preceding the trigger date was used.

2. Coronary procedures

Ischemic heart disease was also identified based on a history of coronary diagnostic or interventional procedures prior to the outpatient visit, including coronary angiography (CAG) and percutaneous coronary intervention (PCI). Procedure codes corresponding to diagnostic angiography and coronary revascularization were included. The most recent qualifying procedure before the trigger date was considered.

3. Biomarker evidence of myocardial injury

Evidence of myocardial injury was defined as a markedly elevated cardiac troponin T concentration (>500 ng/L), based on either the

1 most recent value or the maximum recorded value prior to the
2 outpatient visit.

3 Patients were classified as having ischemic heart disease (IHD_FLAG =
4 1) if any of the above criteria were met prior to the outpatient visit.

5 This composite definition was designed to capture both chronic
6 coronary artery disease and prior acute ischemic events, while
7 minimizing reliance on a single data source.

8 Chronic kidney disease

9 Chronic kidney disease (CKD) was defined based on the most recent

10 estimated glomerular filtration rate (eGFR) available prior to the

11 outpatient visit. CKD stage 3 or higher was defined as eGFR <60

12 mL/min/1.73 m² (CKD60_FLAG), and advanced CKD as eGFR <30

13 mL/min/1.73 m² (CKD30_FLAG). eGFR values were derived from routine

14 laboratory measurements as recorded in the electronic medical record.

15 Implantable cardiac devices (ICD and CRT-D)

16 Presence of an implantable cardiac device was defined based on

17 procedure-related electronic health record signals indicating

- 1 implantation, replacement, revision, or follow-up of a cardiac
- 2 implantable electronic device prior to the outpatient visit.
- 3 Patients were classified as having an implantable cardioverter-
- 4 defibrillator (ICD) if any ICD-related procedure was documented before
- 5 the trigger date (ICD_FLAG = 1).
- 6 Patients were classified as having cardiac resynchronization therapy with defibrillator (CRT-
- 7 D) if procedure codes indicated implantation, replacement, or follow-up of a CRT-D device
- 8 prior to the outpatient visit (CRT_D_FLAG = 1).

Appendix C. Sensitivity analysis evaluating stricter minimum follow-up duration requirements for apparently event-free outpatient visits

Scenario	N	Events (%)	AUROC	AUPRC	Brier
≥550-day follow-up for non-events (main analysis)	199,961	33,627 (16.8%)	0.769 ± 0.003	0.419 ± 0.006	0.118
≥730-day follow-up for non-events (sensitivity analysis)	193,200	33,627 (17.4%)	0.768 ± 0.003	0.427 ± 0.006	0.121

Appendix D. Endpoint-specific sensitivity analyses

Endpoint	AUROC (SD)	AUPRC (SD)	Brier score
2-year composite	0.769 (0.003)	0.418 (0.006)	0.118
2-year hospitalization only	0.751 (0.004)	0.312 (0.014)	0.097
2-year mortality only	0.873 (0.004)	0.335 (0.013)	0.045
1-year composite	0.766 (0.002)	0.309 (0.007)	0.084