American Heart Journal Plus: Cardiology Research and Practice 22 (2022) 100207

Contents lists available at ScienceDirect

American Heart Journal Plus: Cardiology Research and Practice

journal homepage: www.sciencedirect.com/journal/american-heart-journal-plus-cardiology-research-and-practice

Research paper

Development and validation of a model to categorize cardiovascular cause of death using health administrative data

Sagar Patel [a,1], Wade Thompson [b,c,d,e,1], Atul Sivaswamy [c], Anam Khan [c], Laura Ferreira-Legere [c], Douglas S. Lee [c,f,g,h], Husam Abdel-Qadir [c,f,g,h,i], Cynthia Jackevicius [c,f,j], Shaun Goodman [j,k,l], Michael E. Farkouh [h,n], Karen Tu [f,o,p], Moira K. Kapral [c,g], Harindra C. Wijeysundera [c,f,g,m], Derrick Tam [g], Peter C. Austin [c,f], Jiming Fang [c], Dennis T. Ko [c,f,g,m], Jacob A. Udell [c,f,g,h,i,*]

[a] Faculty of Medicine, University of Toronto, Toronto, Canada
[b] Women's College Research Institute, Toronto, Canada
[c] ICES, Toronto, Canada
[d] Research Unit of General Practice, University of Southern Denmark, Odense, Denmark
[e] Department of Anesthesiology, Pharmacology, and Therapeutics, University of British Columbia, Canada
[f] Institute of Health Policy, Management, and Evaluation, University of Toronto, Toronto, Canada
[g] Department of Medicine, Temerty Faculty of Medicine, University of Toronto, Toronto, Canada
[h] Peter Munk Cardiac Centre, University Health Network, Toronto, Canada
[i] Cardiovascular Division, Department of Medicine, Women's College Hospital, Toronto, Canada
[j] Western University of Health Sciences, Pomona, CA, United States of America
[k] Division of Cardiology, St. Michael's Hospital, Toronto, Canada
[l] Canadian VIGOUR Centre, University of Alberta, Edmonton, Canada
[m] Schulich Heart Centre, Sunnybrook Health Sciences Centre, Toronto, Canada
[n] Heart and Stroke/Richard Lewar Centre of Excellence, University of Toronto, Toronto, Canada
[o] North York General Hospital, Department of Family and Community Medicine, University of Toronto, Toronto, Canada
[p] Toronto Western Hospital Family Health Team, University Health Network, Toronto, Canada

ARTICLE INFO

Keywords:
Healthcare outcome assessment
Cohort studies
Databases

ABSTRACT

Study objective: Develop and evaluate a model that uses health administrative data to categorize cardiovascular (CV) cause of death (COD).
Design: Population-based cohort.
Setting: Ontario, Canada.
Participants: Decedents ≥ 40 years with known COD between 2008 and 2015 in the CANHEART cohort, split into derivation (2008 to 2012; n = 363,778) and validation (2013 to 2015; n = 239,672) cohorts.
Main outcome measures: Model performance. COD was categorized as CV or non-CV with ICD-10 codes as the gold standard. We developed a logistic regression model that uses routinely collected healthcare administrative to categorize CV versus non-CV COD. We assessed model discrimination and calibration in the validation cohort.
Results: The strongest predictors for CV COD were history of stroke, history of myocardial infarction, history of heart failure, and CV hospitalization one month before death. In the validation cohort, the c-statistic was 0.80, the sensitivity 0.75 (95 % CI 0.74 to 0.75) and the specificity 0.71 (95 % CI 0.70 to 0.71). In the primary prevention validation sub-cohort, the c-statistic was 0.81, the sensitivity 0.71 (95 % CI 0.70 to 0.71) and the specificity 0.75 (95 % CI 0.75 to 0.75) while in the secondary prevention sub-cohort the c-statistic was 0.74, the sensitivity 0.81 (95 % CI 0.81 to 0.82) and the specificity 0.54 (95 % CI 0.53 to 0.54),
Conclusion: Modelling approaches using health administrative data show potential in categorizing CV COD, though further work is necessary before this approach is employed in clinical studies.

* Corresponding author at: Cardiovascular Division, Peter Munk Cardiac Centre, Toronto General Hospital and Women's College Hospital, University of Toronto, 76 Grenville Street, Toronto, ON M5S 1B1, Canada.
E-mail address: jay.udell@utoronto.ca (J.A. Udell).
[1] Joint first authors.

https://doi.org/10.1016/j.ahjo.2022.100207
Received 22 August 2022; Received in revised form 12 September 2022; Accepted 13 September 2022
Available online 16 September 2022

1. Introduction

Observational studies, and increasingly, pragmatic clinical trials, rely on administrative health data from multiple sources such as electronic health records, physician billing data, and registries. A common outcome of interest in such studies is mortality, in particular, cause-specific mortality. Challenges with use of health administrative data for outcome ascertainment include both the accurate and timely reporting of cause of death (COD) within registries [1–7]. The delay in registering details from death certificates, autopsies, coroner reports and other sources of information due to logistical and administrative barriers is widespread, and has been reported in the United States (US) [1,2], Nordic countries [3–7] and Pacific Island countries [8]. The US National Center for Health Statistics (NCHS) described the process of receiving, processing and editing death records across all 50 states as taking ≥15 months after the calendar year [1,2]. In Sweden, the COD register currently experiences a delay of up to 18 months [3]. Delays in reporting COD is an issue for pragmatic clinical trials and cohort studies, particularly cardiovascular (CV) studies that use CV death as a clinical endpoint, examine cardiovascular safety of medications, or assess the association between CV therapies and CV mortality [9–17]. As very few alternatives to national COD data exist, most studies restrict the outcome ascertainment to all-cause mortality or wait several years to retrospectively report outcomes [10]. A valid prediction model that can discriminate between CV and non-CV cause of death, would be useful to facilitate timely research and outcome reporting for projects involving health administrative datasets, as well as timely tracking of population health. The objective of this study was to develop and validate a model to estimate the probability that a given death was due to CV causes using detailed patient characteristics and traditional cardiac risk factors.

2. Methods

2.1. Data source

We conducted this study using Cardiovascular Health in Ambulatory Care Research Team (CANHEART) cohort [18,19], a registry-based cohort which contains data on almost the entire Ontario community-dwelling adult population of 10.9 million individuals. This cohort is housed at ICES, an independent, non-profit research institute whose legal status under Ontario's health information privacy law allows it to collect and analyze health care and demographic data, without consent, for health system evaluation and improvement. The Service Ontario Office of the Registrar General of Ontario Vital Statistics Database (ORGD) is an annual dataset that contains all the deaths recorded in Ontario since 1990. The ORGD contains specific direct patient identifiers such as name, date of birth, and resident postal code; as well as information on date of death and cause of death. Since unique patient identifiers that are assigned to each resident of Ontario are not collected in the ORGD, we could not use it as a direct source of patient identification in order to link the death data to other databases. Hence, deaths were linked to other databases using both deterministic and probabilistic linkage. The datasets in the CANHEART cohort include individual-level information on socio-demographic characteristics (Registered Persons Database; Immigration, Refugees and Citizenship Canada Permanent Resident Database), past medical history (via Canadian Institute for Health Information [CIHI] Discharge Abstract Database [DAD]), medications on those aged 65 and older (Ontario Drug Benefit Database), and health care services use (CIHI-DAD, National Ambulatory Care Reporting System [NACRS]), and were linked using unique encoded identifiers and analyzed at ICES. Follow-up for clinical events was obtained from linkage to CIHI-DAD. Disease and procedure definitions used were standard CANHEART definitions [18].

2.2. Study population

All residents in Ontario, Canada aged 40 to 105 years and eligible for the province's Ontario Health Insurance Plan (OHIP) as of January 1, 2008 were considered for inclusion in our cohort. The sample was split into a model derivation cohort (individuals who died between January 1st, 2008 and December 31st, 2012 (eFig. 1) and a temporally distinct validation cohort (individuals who died between January 1st, 2013 and December 31st, 2015 (eFig. 2). In addition, all deaths with missing information on cause of death, as noted in ORGD, were excluded.

2.3. Ascertainment of cause of death

The ORGD uses International Classification of Disease Tenth Edition (ICD-10) codes to categorize the underlying cause of death. We considered the cause of death to be cardiovascular if the ICD-10 code was any of I00-I78 (“diseases of the circulatory system”) [20]. All other individuals in the cohort were classified as having died of non-cardiovascular causes. The ORGD provides COD information obtained from death certificates, which are completed by a treating physician or coroner.

2.4. Statistical analysis

2.4.1. Development of the model to ascertain cause of death in administrative data

We included the following characteristics and risk factors in our model (see eTable 1 for definitions): age, sex, ethnicity (based on surname algorithm), rural location, low income (based on neighbourhood income quintiles), co-morbidities, history of CV procedures, hospitalization within 1 month of death, and healthcare utilization in the 2 years prior to death. Patient baseline characteristics were presented using descriptive statistics. Those that died of CV causes were compared to those who died of non-CV causes. We compared baseline characteristics in the two groups using the standardized mean difference (SMD).

Logistic regression was used to develop the prediction model. We first conducted univariate analyses for the association between each variable and CV death, with a conservative p-value of <0.25, to identify variables that could be included in our final model. With this refined group of variables, we then used backwards elimination to develop our final model, which only included variables with $p < 0.05$ for the association between the variable and CV death. Variables deemed clinically relevant a priori were included in the model irrespective of statistical significance. We captured health system utilization in the two years before death as a baseline characteristic; however, we elected not to include these variables in the final prediction model as this would limit the utility of our model to jurisdictions with different patterns of health care use.

2.4.2. Assessment of model performance

Model performance was assessed using discrimination and calibration measures and using the ORGD classification of COD as the gold standard. Discrimination was measured using the c-statistic (area under the Receiver Operating characteristic [ROC] curve), with a value over 0.70 indicative of good discrimination [21]. Calibration was assessed by comparing the agreement between observed probabilities and predicted probability of CV death across the ten deciles of predicted risk of CV death.

2.4.3. Cut-off value for assigning cardiovascular disease as the cause of death

The predicted probability of CV death was computed for each individual in the development cohort based on the final model. We calculated the sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and Youden's Index, ranging from a predicted probability of 0.05 to 0.95 in increments of 0.05 [22], using

the actual cause of death in the ORGD as the gold standard. The highest value of Youden's Index was set as the ideal cut-off threshold for the probability of cardiovascular death (i.e., maximized the correct classification of individuals as having died from CV causes).

2.4.4. Validation cohort

We applied our final prediction model to individuals in the validation cohort, and assessed the c-statistic, sensitivity, specificity, PPV, and NPV using the optimal cut-off of 30 %. Calibration was assessed the same way it was for the internal validation cohort. We also calculated the Brier score (ranging from 0 to 1) to examine model performance for the validation cohort, where a Brier score close to 0 denotes better accuracy. Within the validation cohort, we defined and assessed model performance in the following cohorts which reflect inclusion criteria of populations often studied in cardiovascular clinical trials or cohort studies: i) primary prevention, defined as those without CV disease, ii) secondary prevention, consisting of those with a history of CV disease, iii) prior percutaneous coronary intervention (PCI) or coronary artery bypass graft (CABG), but no prior acute myocardial infarction (AMI), iv) prior AMI but without either stroke or peripheral artery disease (PAD), v) prior AMI and at least one of stroke or PAD, vi) history of AMI, stroke, and PAD, vii) congestive heart failure (CHF) with at least one of AMI, PCI, or CABG.

All data were analyzed at ICES using SAS version 9.4 (SAS Institute, Cary, NC). The use of data in this project was authorized under section 45 of Ontario's Personal Health Information Protection Act, which does not require review by a Research Ethics Board.

3. Results

3.1. Baseline characteristics

Our model derivation cohort consisted of 362,778 patients who died between January 1, 2008 and December 31, 2012 (eFig. 1); 258,188 (71.2 %) were non-CV deaths and 104,590 (28.8 %) were CV deaths. Baseline patient characteristics are shown in Table 1. For our validation cohort, we identified 239,672 individuals who died between January 1, 2013 to December 31, 2015 (eFig. 2), of which 65,806 (27.5 %) were CV deaths. The primary prevention cohort consisted of 180,061 individuals and the secondary prevention cohort of 59,611 with 23.7 % and 33.8 % of individuals dying of cardiovascular causes respectively (eTable 2).

3.2. Development cohort and model performance

The univariate associations between predictors and CV death are in eTable 3. The final set of predictors is in Table 2. Positive predictors of cardiovascular death included: increasing age, male sex, rural location, low-income neighbourhood, South Asian ethnicity (based on surname), history of MI, history of stroke, history of CHF, history of PCI or CABG, history of TIA, history of arrhythmia, a CV hospitalization in the 1 month before death, or an eye-related hospitalization in the 1 month before death (Table 2). Negative predictors included: cancer, HIV, diabetes, chronic dialysis, and hospitalizations for various non-CV issues in the 1 month before death (see Table 2). The c-statistic for the predictive model in the derivation sample was 0.831 (eFig. 3). The observed versus predicted probability of CV death by risk decile in the development cohort is in Fig. 1, and a calibration plot is in eFig. 5.

3.3. Determining cut-off for model

The sensitivity of our model decreased, and specificity increased, as the cut-off value of the predicted probability for CV death increased (eTable 4 and eFig. 4). The highest Youden's J statistic value, 0.52, for the algorithm was seen at a cut off value where the probability of CV death was 30 %. This indicated that a cut-off of 30 % predicted probability of CV death would maximize correct classification of CV death. At

Table 1
Baseline characteristics of the cohort across cause of death in the derivation cohort (2008 to 2012).

		Non-CV death n (%) N = 258,188	CV death n (%) N = 104,590	Standardized difference
Socio-demographic characteristics				
Age at time of death, years	Mean ± SD	74.9 ± 13.3	78.8 ± 12.4	0.31
	Median (IQR)	77 (66–85)	82 (72–88)	0.31
Age group (time of death), years	40–49	12,359 (4.8 %)	2572 (2.5 %)	0.12
	50–59	27,283 (10.6 %)	7122 (6.8 %)	0.13
	60–69	43,036 (16.7 %)	12,953 (12.4 %)	0.12
	70–79	63,536 (24.6 %)	22,508 (21.5 %)	0.07
	80+	111,974 (43.4 %)	59,435 (56.8 %)	0.27
Female		122,967 (47.6 %)	48,654 (46.5 %)	0.02
Low income[a]		112,701 (43.7 %)	46,863 (44.8 %)	0.02
Ethnic Group	Chinese	5180 (2.0 %)	1624 (1.6 %)	0.03
	South Asian	3005 (1.2 %)	1433 (1.4 %)	0.02
	General	250,003 (96.8 %)	101,533 (97.1 %)	0.01
Rural status[b]		37,411 (14.5 %)	15,994 (15.3 %)	0.02
	Missing	158 (0.1 %)	58 (0.1 %)	0.02
Healthcare utilization performed in the 2 years before death				
Cholesterol screening[c]		137,649 (53.3 %)	58,421 (55.9 %)	0.05
Diabetes screening[d]		198,052 (76.7 %)	80,676 (77.1 %)	0.01
Electrocardiogram		229,392 (88.8 %)	89,672 (85.7 %)	0.09
Echocardiogram		109,500 (42.4 %)	55,345 (52.9 %)	0.21
Stress test		24,255 (9.4 %)	11,874 (11.4 %)	0.06
Visit to cardiologist		52,833 (20.5 %)	33,796 (32.3 %)	0.27
Visit to family doctor		252,048 (97.6 %)	99,899 (95.5 %)	0.12
Visit to specialist		235,639 (91.3 %)	87,698 (83.8 %)	0.23
Disease history and cardiovascular procedures				
Acute myocardial infarction		39,892 (15.5 %)	33,941 (32.5 %)	0.41
CABG		13,393 (5.2 %)	10,654 (10.2 %)	0.19
Heart failure		54,382 (21.1 %)	42,181 (40.3 %)	0.43
Ischemic Heart Disease		75,857 (29.4 %)	53,539 (51.2 %)	0.46
Peripheral Arterial Disease		13,481 (5.2 %)	8966 (8.6 %)	0.13
PCI		12,209 (4.7 %)	9041 (8.6 %)	0.16
PCI or CABG		23,361 (9.0 %)	17,486 (16.7 %)	0.23

(continued on next page)

Table 1 (*continued*)

	Non-CV death n (%)	CV death n (%)	Standardized difference
	N = 258,188	N = 104,590	
Stroke	23,452 (9.1 %)	26,034 (24.9 %)	0.43
Transient Ischemic Attack	8349 (3.2 %)	5735 (5.5 %)	0.11
Cancer	151,264 (58.6 %)	21,378 (20.4 %)	0.85
COPD	103,284 (40.0 %)	40,485 (38.7 %)	0.03
Diabetes	89,296 (34.6 %)	38,622 (36.9 %)	0.05
Endarterectomy	2971 (1.2 %)	2111 (2.0 %)	0.07
Hypertension	184,224 (71.4 %)	86,534 (82.7 %)	0.27
Human Immunodeficiency Virus	795 (0.3 %)	97 (0.1 %)	0.05
Rheumatoid arthritis	7904 (3.1 %)	3343 (3.2 %)	0.01
Atrial Fibrillation	37,918 (14.7 %)	27,124 (25.9 %)	0.28
Arrhythmia	78,742 (30.5 %)	52,771 (50.5 %)	0.42
Chronic dialysis	6057 (2.3 %)	2639 (2.5 %)	0.01
Main diagnosis for hospitalization one month prior to death			
Blood-related	1539 (0.6 %)	271 (0.3 %)	0.05
Cardiovascular-related	12,323 (4.8 %)	30,543 (29.2 %)	0.69
Digestive-related	13,011 (5.0 %)	1681 (1.6 %)	0.19
Endocrine-related	3710 (1.4 %)	977 (0.9 %)	0.05
Eye-related	286 (0.1 %)	198 (0.2 %)	0.02
Infection-related	11,074 (4.3 %)	1476 (1.4 %)	0.17
Injury-related	8525 (3.3 %)	1808 (1.7 %)	0.1
Mental health-related	1432 (0.6 %)	361 (0.3 %)	0.03
Muscle-related	1135 (0.4 %)	346 (0.3 %)	0.02
Neoplasm-related	27,423 (10.6 %)	573 (0.5 %)	0.45
Nervous system-related	2207 (0.9 %)	828 (0.8 %)	0.01
Respiratory-related	25,282 (9.8 %)	4345 (4.2 %)	0.22
Skin-related	655 (0.3 %)	235 (0.2 %)	0.01

CABG = coronary artery bypass graft; COPD = chronic obstructive pulmonary disease; CV = cardiovascular; ECG = electrocardiogram; ECHO = echocardiogram; PCI = percutaneous coronary intervention.
Patients included almost all Ontarians from 40 to 105 years of age with recorded all-cause mortality between Jan 1, 2008–Dec 31, 2012. Cardiovascular death was categorized based on ICD-10 codes I00 to I78.
[a] Low income is defined as being in bottom two quintiles of provincial neighbourhood income.
[b] Rural status defined as living in a community of <10,000 individuals.
[c] Cholesterol screening = Ontario Health Insurance Plan (OHIP) fee codes L055, L117, and L243 on same day in 2 years before death (yes/no).
[d] Diabetes screening = OHIP fee codes G498, L104, L111 or L093 in 2 years before death (yes/no).

this cut-off, the model had a sensitivity of 0.77 (95 % CI 0.76 to 0.77), specificity of 0.75 (95 % CI 0.74 to 0.75), PPV of 0.55 (95 % CI 0.55 to 0.55), and NPV of 0.89 (95 % CI 0.89 to 0.89).

3.4. Validation cohort

Model performance in the validation cohort is presented in Table 3. The c-statistic in the overall cohort was 0.80 and the Brier score was 0.15. The c-statistics in the primary prevention and secondary prevention cohort were 0.81 and 0.74, respectively. In the overall validation cohort, using a cut-off of 30 % probability to classify CV death, the model had a sensitivity of 0.75 (95 % CI 0.74 to 0.75), specificity 0.71 (95 % CI 0.70 to 0.71), PPV 0.49 (95 % CI 0.49 to 0.49), and NPV 0.88 (95 % CI 0.88 to 0.88). The primary prevention cohort model had a sensitivity of 0.71 (95 % CI 0.70 to 0.71), specificity 0.75 (95 % CI 0.75 to 0.75), PPV 0.47 (95 % CI 0.47 to 30.47) and NPV 0.89 (95 % CI 0.89 to 0.89). The secondary prevention cohort model had a sensitivity of 0.81 (95 % CI 0.81 to 0.82), specificity 0.54 (95 % CI 0.53 to 0.54), PPV 0.53 (95 % CI 0.52 to 0.53), and NPV 0.82 (95 % CI 0.81 to 0.82). The findings from the additional secondary prevention cohorts can be found in Table 3. The observed versus predicted probability of CV death in the validation cohort (overall and for the primary and secondary prevention subgroups) is presented in Fig. 2, and a calibration plot is displayed in eFig. 6.

Table 2
Odds ratios and 95 % confidence intervals for the fully adjusted model predicting cardiovascular mortality in the derivation cohort (n = 362,778) between 2008 and 2012.

Characteristic	Odds ratio (95 % CI)
Socio-demographics	
Age at time of death, per year increase	1.01 (1.01,1.01)
Male sex	1.20 (1.18,1.22)
Rurality[a]	
Non-rural (ref)	–
Rural	1.11 (1.08,1.13)
Missing	1.02 (0.70,1.48)
Income[b]	
Medium/high income (ref)	–
Low income	1.04 (1.02,1.06)
Missing	0.94 (0.83,1.05)
Ethnicity	
Chinese (ref)	–
General	1.19 (1.11,1.27)
South Asian	1.18 (1.07,1.3)
Past medical history	
Stroke	2.16 (2.11,2.21)
Acute Myocardial Infarction	1.48 (1.45,1.52)
Heart failure	1.44 (1.41,1.47)
PCI or CABG	1.19 (1.15,1.22)
Peripheral Arterial Disease	1.16 (1.12,1.2)
Transient Ischemic Attack	1.09 (1.04,1.13)
Hypertension	1.29 (1.26,1.32)
Arrythmia	1.50 (1.47,1.53)
Rheumatoid arthritis	0.94 (0.9,0.99)
Asthma	0.90 (0.88,0.93)
COPD	0.83 (0.81,0.85)
Diabetes	0.77 (0.76,0.79)
Chronic dialysis	0.63 (0.59,0.66)
Human Immunodeficiency Virus	0.4 (0.31,0.5)
Cancer	0.21 (0.21,0.21)
Hospitalization within 1 month before death	
Cardiovascular	3.94 (3.83,4.04)
Eye	1.43 (1.14,1.79)
Nervous system	0.63 (0.57,0.68)
Skin	0.58 (0.5,0.69)
Musculoskeletal	0.53 (0.47,0.61)
Blood related disorders	0.51 (0.44,0.59)
Endocrine disorders	0.49 (0.45,0.53)
Mental health	0.41 (0.36,0.47)
Injuries	0.30 (0.28,0.32)
Respiratory	0.27 (0.26,0.28)
Digestive	0.24 (0.23,0.25)
Infectious disease	0.22 (0.21,0.23)
Neoplasms	0.13 (0.12,0.14)

CABG = coronary artery bypass graft; CI = confidence interval; COPD = chronic obstructive pulmonary disease; PCI = percutaneous coronary intervention.
Final predictors obtained via backwards elimination and a threshold of $p < 0.05$.
[a] Rural status defined as living in a community of <10,000 individuals.
[b] Low income is defined as being in income quintile 1 or 2.

4. Discussion

4.1. Summary of findings

We used the CANHEART cohort to develop and validate a model that

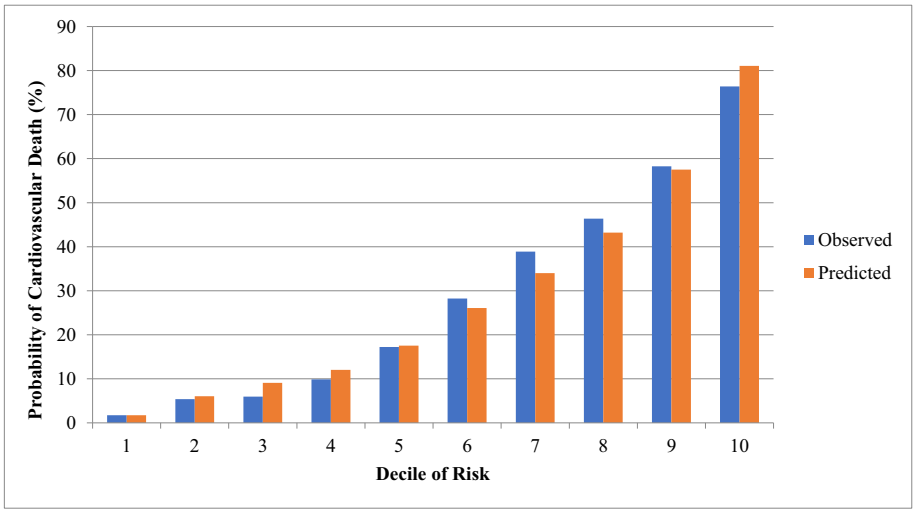

Fig. 1. Observed and predicted probability of cardiovascular death in the derivation cohort (2008 to 2012) across deciles of risk for cardiovascular-related death. Comparison of predicted deaths using our CV prediction model, to observed deaths captured using Registrar General of Ontario Vital Statistics Database.

Table 3
Sensitivity, specificity, PPV and NPV, and C-statistic in the validation cohort (2013 to 2015).

Cohort	Sensitivity	Specificity	PPV	NPV	C-statistic	Brier score
Overall cohort	0.75 (0.74–0.75)	0.71 (0.70–0.71)	0.49 (0.49–0.49)	0.88 (0.88–0.88)	0.80	0.15
Primary cohort	0.71 (0.70–0.71)	0.75 (0.75–0.75)	0.47 (0.47–0.47)	0.89 (0.89–0.89)	0.81	0.14
Secondary cohort	0.81 (0.81–0.82)	0.54 (0.53–0.54)	0.53 (0.52–0.53)	0.82 (0.81–0.82)	0.74	0.20
(PCI or CABG) and no AMI Cohort	0.78 (0.76–0.79)	0.61 (0.60–0.62)	0.54 (0.52–0.55)	0.82 (0.81–0.83)	0.76	0.19
AMI and no (Stroke or PAD) Cohort	0.82 (0.81–0.83)	0.55 (0.55–0.56)	0.55 (0.54–0.55)	0.83 (0.82–0.83)	0.76	0.20
AMI and (Stroke or PAD) Cohort	0.91 (0.90–0.93)	0.33 (0.31–0.35)	0.52 (0.50–0.53)	0.83 (0.80–0.85)	0.72	0.22
AMI and Stroke and PAD Cohort	0.99 (0.96–1.00)	0.15 (0.10–0.21)	0.52 (0.47–0.58)	0.93 (0.77–0.99)	0.67	0.25
CHF and (AMI or PCI or CABG) Cohort	0.85 (0.84–0.86)	0.42 (0.40–0.43)	0.56 (0.55–0.57)	0.76 (0.74–0.77)	0.71	0.22

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CHF = congestive heart failure; NPV = negative predictive value; PCI = percutaneous coronary intervention; PPV = positive predictive value.
Using the final prediction model and a cut-off of 30 % probability of CV death to classify a cause of death as CV, we calculated the sensitivity, specificity, PPV, NPV, and C-statistic for our model using a gold standard of cause of death captured via the Registrar General of Ontario Vital Statistics Database.

could categorize deaths as being from CV causes using routinely collected healthcare administrative data, with the actual cause of death as noted in the ORGD as the gold standard. Important predictors of CV versus non-CV death included distant and recent antecedent CV events, risk factors, and co-morbidities. When applied to a validation cohort, our model displayed modest performance in terms of sensitivity, specificity, PPV, and NPV, and calibration. There was discordance between observed and predicted CV COD in the validation cohort, suggesting further work is required before this model is used to categorize CV COD in observational studies or pragmatic trials.

4.2. Comparison to existing literature

The delay in COD information appearing in registries and databases is in part due to the time necessary for stakeholders to collect, verify, code and transfer the data [8]. Currently, registry-based trials that use national death data do not have many alternatives around the delayed COD data. Wong et al. [10] and Maynard et al. [9] have described the delay of COD data in the US National Death Index and Washington State death registry, respectively, as a major limitation to studies relying on COD data. There have been few efforts to use routinely collected health administrative data to accurately categorize or predict CV mortality. The characteristics we found to be predictive of CV death (e.g., CV co-morbidities, CV hospitalization in month before death) were as expected, given those with existing CVD are at elevated risk for future and fatal CV events [23]. An algorithm to identify CV death using health administrative data was developed in Canada and the UK using data from 20,000 individuals, with an overall sensitivity of 65 % and a sensitivity of 63 % in Ontario (the authors concluded the algorithm had moderate validity) [24]. This study classified CV death by using either in-hospital death with CV diagnosis or out-of-hospital death excluding cancer or trauma prior to death date. Our approach incorporating previous diagnoses and healthcare utilization before death yielded improved validity compared to this study. Li et al. developed a model to predict future CV-specific mortality based on questionnaire data from a prospective cohort study of 2359 people in Taiwan, investigating socio-demographics (age, sex, income, education) and clinical data (vitals, lab values, co-morbidities, medication use) and linking to Taiwan's National Registry of Death [25]. These authors found that age, history of CVD, an ankle-brachial index ≤ 0.9, and CV medication use were all predictive of CV death, though sex was not. The c-statistic ranged from 0.88 to 0.98 depending on the time horizon for the CV death. Sherazi et al. used different machine learning approaches to predict mortality in the 1-year post-acute coronary syndrome (ACS) from a registry in Korea ($n = 10{,}813$), finding that predictors varied depending on which approach was used though common ones included aspirin use, diuretic use, age $>$ 76 years, or high LDL [26]. The c-statistic ranged from 0.81 to 0.90. The aim of both these studies was to predict CV mortality rather than categorize CV COD. Further, they both involved standardized data collection approaches. Our novel contribution beyond these efforts includes our use of data that are generally routinely captured across various regions/jurisdictions, which potentially enables external validation and applicability outside of Ontario.

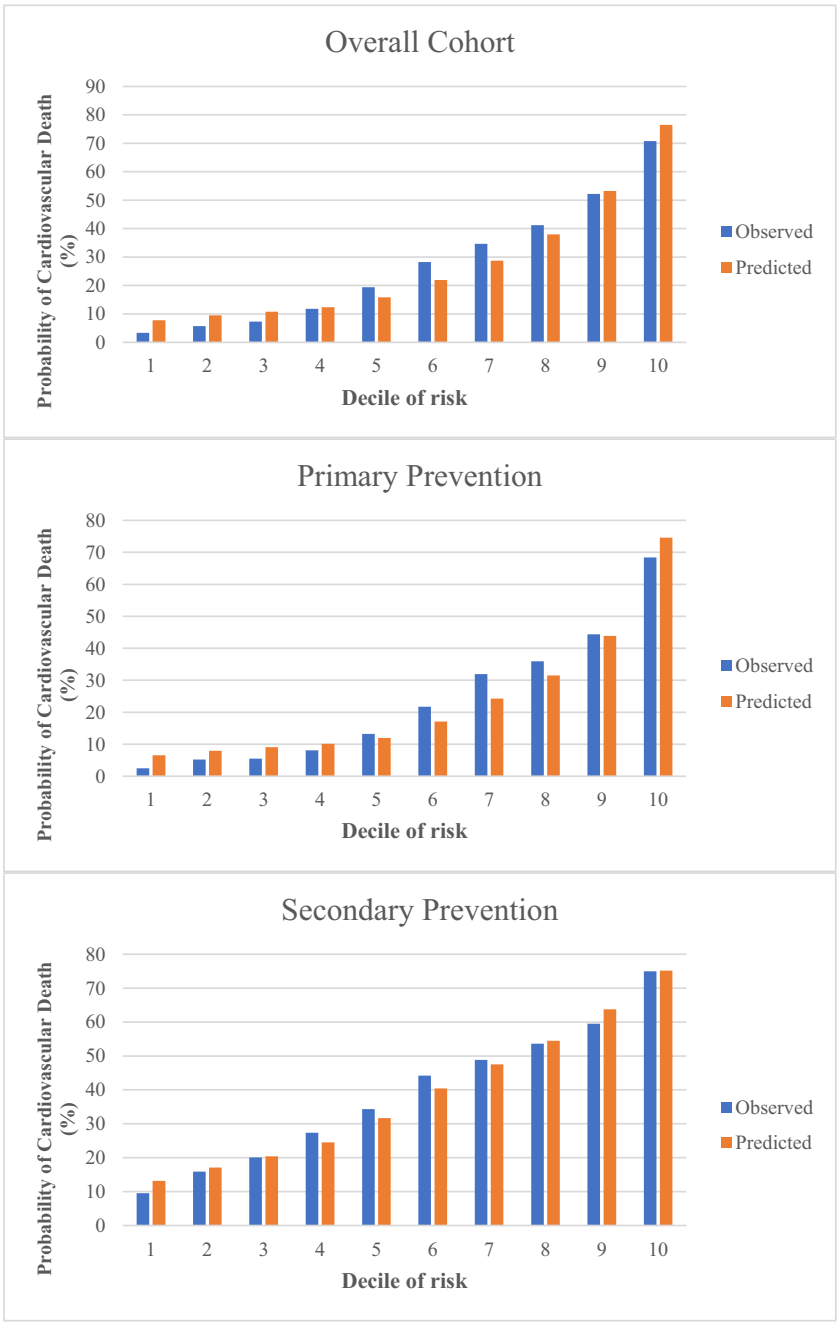

Fig. 2. Observed and predicted probability of cardiovascular death in the validation cohort (2013 to 2015) across deciles of predicted risk for cardiovascular-related death, for overall cohort and subgroups of primary prevention and secondary prevention.
Comparison of predicted deaths using our CV prediction model, to observed deaths captured using Registrar General of Ontario Vital Statistics Database.

4.3. Implications and future directions

Not having readily available COD data from health administrative databases presents a challenge to providing timely and contemporary investigation of new guidelines, real-world safety and effectiveness of drugs, or investigating outcomes of interest in pragmatic clinical trials relying on health administrative data. Statistics Canada and the US NCHS are both aware of this issue and are looking at workarounds to these delays [1,2]. For example, NCHS launched a National Death Index (NDI) early release pilot program in 2019 [27]. Our work highlights the potential for developing predictive models to address the current lack of timely COD data in healthcare administrative databases. However, further work is necessary before using this approach in registry-based trials or observational studies. Refining the modelling approach to improve accuracy will enhance its usefulness. Determining a threshold for how accurate such a model must be in this context is also a consideration for future work. Given the performance of the model appears to differ in the primary prevention population compared to secondary prevention, future studies could develop separate models in these different populations a priori. Future work could also employ different development approaches, such as machine learning or multiple imputation, which may enhance the ability to identify important predictors of CV death. Machine learning approaches have shown promise in predicting mortality in cancer [28], and have been applied to predicting post-acute coronary syndrome mortality [26]. It will also be important to evaluate a predictive modelling approach using different datasets to

examine the model's reproducibility, reliability, and accuracy in different healthcare systems.

4.4. Limitations

To develop the prediction model using the derivation cohort, we combined the primary prevention and secondary prevention population into one cohort. However, in our validation process, we found differences in model performance between the primary and secondary prevention cohort. Thus, it may have been appropriate to develop separate models for primary and secondary prevention. We were unable to include information on BMI or self-reported behavior and lifestyle information such as smoking, which is not well-captured at an individual patient level in the CANHEART cohort. Further, in developing our model and balancing external validity and applicability outside Ontario, we did not use all available data sources (e.g., prescription drug data among persons $\geq$65 years). The model was developed using healthcare administrative data from the CANHEART cohort in Ontario, Canada and it is unclear whether our model would be generalizable in other jurisdictions and using their respective datasets. This requires further study, though we expect our model to be adaptable to contexts which have similar data in terms of content completeness, and quality. Our performance data suggest that the model may still misclassify some deaths, highlighting the importance of future work to improve model performance. Given our validation and derivation cohorts were not from the same time period, it is possible that temporal trends in CV death or population characteristics may have influenced our results. However, we are not aware of any trends that would have such an influence. Finally, while we considered the COD in the ORGD (based on the death certificate) as the gold standard, the accuracy of COD in ORGD has not been examined and we did not verify the accuracy of the ORGD data. Indeed, the accuracy of COD data in death certificates in general has been questioned [29].

5. Conclusion

We developed a model that used routinely collected health administrative data to categorize CV COD. Strong predictors of CV COD included cardiovascular co-morbidities and cardiovascular hospitalizations one month before death. In a validation cohort, our model showed modest performance in categorizing CV death. Our results highlight the potential for a modelling approach to categorize CV COD from health administrative data, though further work is necessary to improve the performance of such models before they could be routinely implemented for observational research and pragmatic trials.

CRediT authorship contribution statement

Conceptualization and Methodology: all authors.
Formal Analysis: AS, AK, PA, SP, JAU.
Investigation: all authors.
Original draft: SP, WT, JAU.
Review and editing: all authors.
Supervision: JAU.
Project administration: LFL, JAU.
Funding acquisition: JAU.

Declaration of competing interest

Dr. Udell is supported by a Heart and Stroke Foundation National New Investigator-Ontario Clinician Scientist Award; Ontario Ministry of Research, Innovation and Science Early Researcher Award; grants from AstraZeneca, Novartis, and Sanofi. Dr. Udell reports receiving personal fees for consulting for or honoraria from Amgen, AstraZeneca, Boehringer-Ingelheim, Janssen, Merck, Novartis and Sanofi. Dr. J. Tu was supported by a Tier 1 Canada Research Chair in Health Services Research and an Eaton Scholar award from the Department of Medicine, University of Toronto. Dr. Austin is supported by a Mid-Career Investigator Award from the Heart and Stroke Foundation. Dr. Lee is the Ted Rogers Chair in Heart Function Outcomes, University Health Network, University of Toronto. Dr. Farkouh is the Peter Munk Chair in Multinational Clinical Trials at Peter Munk Cardiac Centre, University Health Network, University of Toronto. Dr. K Tu receives a Research Scholar Award from the Department of Family and Community Medicine, University of Toronto. Dr. Goodman receives research grant support (e.g., steering committee or data and safety monitoring committee) and/or speaker/consulting honoraria (e.g., advisory boards) from: Amgen, Anthos Therapeutics, AstraZeneca, Bayer Canada, Boehringer Ingelheim, Bristol Myers Squibb, CSL Behring, Daiichi-Sankyo/American Regent, Eli Lilly and Company, Esperion, Ferring Pharmaceuticals, HLS Therapeutics, JAMP Pharma, Merck, Novartis, Novo Nordisk A/C, Pendopharm/Pharmascience, Pfizer, Regeneron, Sanofi, Servier, Valeo Pharma; and salary support/honoraria from the Heart and Stroke Foundation of Ontario/University of Toronto (Polo) Chair, Canadian Heart Research Centre and MD Primer, Canadian VIGOUR Centre, Cleveland Clinic Coordinating Centre for Clinical Research, Duke Clinical Research Institute, New York University Clinical Coordinating Centre, PERFUSE Research Institute, TIMI Study Group (Brigham Health). Dr. Kapral holds the Lillian Love Chair in Women's Health at the University Health Network/University of Toronto.

Acknowledgements

Dr. Jack V Tu is posthumously acknowledged for his contribution to the present work.

Funding/support

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). This study was supported by operating grants from the Institute for Circulatory and Respiratory Health–Canadian Institutes of Health Research (CIHR) Chronic Diseases Team (grant no. TCA 118349), a CIHR–Foundation Scheme grant (grant no. FDN 143313), and a CIHR Strategy for Patient Oriented Research Innovative Clinical Trial Multi-Year Grant (grant no. MYG 151211). Parts of this material are based on data and information compiled and provided by the MOH, Canadian Institute for Health Information (CIHI), Cancer Care Ontario (CCO), Ontario Registrar General (ORG) information on deaths, the original source of which is Service Ontario. The analyses, conclusions, opinions and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred. Parts or whole of this material are based on data and/or information compiled and provided by Immigration, Refugees and Citizenship Canada (IRCC) current to December 31, 2015. However, the analyses, conclusions, opinions and statements expressed in the material are those of the author(s), and not necessarily those of IRCC. We thank IQVIA Solutions Canada Inc. for use of their Drug Information Database.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ahjo.2022.100207.

References

[1] M. Siri, C. DL, Statistics CoN, Council NR, Education DoBaSSa, Vital Statistics: Summary of a Workshop, The National Academies Press, 2009.
[2] Statistics NCfH, NDI early release pilot program, Centers for Disease Control and Prevention, 2020. https://www.cdc.gov/nchs/ndi/ndi_early_release.htm.

[3] H.L. Brooke, M. Talback, J. Hornblad, et al., The swedish cause of death register, Eur. J. Epidemiol. 32 (9) (Sep 2017) 765–773, https://doi.org/10.1007/s10654-017-0316-1.
[4] N. Hammar, L. Alfredsson, M. Rosen, C.L. Spetz, T. Kahan, A.S. Ysberg, A national record linkage to study acute myocardial infarction incidence and case fatality in Sweden, Int. J. Epidemiol. 30 (Suppl 1) (Oct 2001) S30–S34.
[5] K. Helweg-Larsen, The Danish register of causes of death, Scand. J. Public Health 39 (7 Suppl) (Jul 2011) 26–29, https://doi.org/10.1177/1403494811399958.
[6] J.F. Ludvigsson, S.E. Haberg, G.P. Knudsen, et al., Ethical aspects of registry-based research in the Nordic countries, Clin. Epidemiol. 7 (2015) 491–508, https://doi.org/10.2147/CLEP.S90589.
[7] P. Nordstrom, Y. Gustafson, K. Michaelsson, A. Nordstrom, Length of hospital stay after hip fracture and short term risk of death after discharge: a total cohort study in Sweden, BMJ 350 (Feb 20 2015), h696, https://doi.org/10.1136/bmj.h696.
[8] K.L. Carter, C. Rao, A.D. Lopez, R. Taylor, Mortality and cause-of-death reporting and analysis systems in seven Pacific Island countries, BMC Public Health 12 (Jun 13 2012) 436, https://doi.org/10.1186/1471-2458-12-436.
[9] C. Maynard, E. Lowy, S.D. Fihn, M. McDonell, Cause of death in Washington state veterans hospitalized with acute coronary syndromes in the veterans health administration, Popul. Health Metrics 6 (Jul 23 2008) 3, https://doi.org/10.1186/1478-7954-6-3.
[10] M.D. Wong, A.K. Chung, W.J. Boscardin, The contribution of specific causes of death to sex differences in mortality, Public Health Rep. 121 (6) (Nov-Dec 2006) 746–754, https://doi.org/10.1177/003335490612100615, 09/14/2020.
[11] J.J. Wang, G. Liew, T.Y. Wong, et al., Retinal vascular calibre and the risk of coronary heart disease-related death, Heart 92 (11) (Nov 2006) 1583–1587, https://doi.org/10.1136/hrt.2006.090522.
[12] T.Y. Wong, R. Klein, F.J. Nieto, et al., Retinal microvascular abnormalities and 10-year cardiovascular mortality: a population-based case-control study, Ophthalmology 110 (5) (May 2003) 933–940, https://doi.org/10.1016/S0161-6420(03)00084-8.
[13] J. Amin, M.G. Law, M. Bartlett, J.M. Kaldor, G.J. Dore, Causes of death after diagnosis of hepatitis B or hepatitis C infection: a large community-based linkage study, Lancet 368 (9539) (2006) 938–945, https://doi.org/10.1016/S0140-6736(06)69374-4.
[14] G. Palmer, M.A. Herbert, S.L. Prince, Coronary Artery Revascularization (CARE) registry: an observational study of on-pump and off-pump coronary artery revascularization, Ann. Thorac. Surg. 83 (3) (Mar 2007) 986–991, https://doi.org/10.1016/j.athoracsur.2006.10.057, discussion 991-2.
[15] A. Klein, K. Lee, A. Gera, T.A. Ports, A.D. Michaels, Long-term mortality, cause of death, and temporal trends in complications after percutaneous aortic balloon valvuloplasty for calcific aortic stenosis, J. Interv. Cardiol. 19 (3) (Jun 2006) 269–275, https://doi.org/10.1111/j.1540-8183.2006.00142.x.
[16] A.B. Patel, J.B. Kostis, A.C. Wilson, M.L. Shea, S.L. Pressel, B.R. Davis, Long-term fatal outcomes in subjects with stroke or transient ischemic attack: fourteen-year follow-up of the systolic hypertension in the elderly program, Stroke 39 (4) (Apr 2008) 1084–1089, https://doi.org/10.1161/STROKEAHA.107.500777.
[17] S.L. Murphy, J. Xu, K.D. Kochanek, S.C. Curtin, E. Arias, Deaths: final data for 2015, Natl. Vital Stat. Rep. 66 (6) (Nov 2017) 1–75.
[18] J.V. Tu, A. Chu, L.R. Donovan, et al., The cardiovascular health in ambulatory care research team (CANHEART): using big data to measure and improve cardiovascular health and healthcare services, Circ. Cardiovasc. Qual. Outcomes 8 (2) (Mar 2015) 204–212, https://doi.org/10.1161/CIRCOUTCOMES.114.001416.
[19] M. Chiu, M. Lebenbaum, K. Lam, Describing the linkages of the immigration, refugees and citizenship Canada permanent resident data and vital statistics death registry to Ontario's administrative health database, BMC Med. Inform. Decis. Mak. 16 (1) (Oct 21 2016) 135, https://doi.org/10.1186/s12911-016-0375-3.
[20] L. Geran, P. Tully, P. Wood, B. Thomas, Comparability of ICD-9 and ICD-10 for Mortality Statistics in Canada, Statistics Canada, Ottawa, 2005. Catalogue no 84-584-XIE.
[21] P.S. Kamath, R.H. Wiesner, M. Malinchoc, et al., A model to predict survival in patients with end-stage liver disease, Hepatology 33 (2) (Feb 2001) 464–470, https://doi.org/10.1053/jhep.2001.22172.
[22] J.R. Lambert, E.K. Lilly, I. Lipkovich, A Macro For Getting More Out Of Your ROC Curve SAS Global Forum, Paper 231, 2008.
[23] V. SS, Heart disease and stroke statistics-2020 update: a report from the American Heart Association, Circulation 141 (9) (2020), https://doi.org/10.1161/CIR.0000000000000757, 03/03/2020.
[24] L.M. Lix, S. Sobhan, A. St-Jean, et al., Validity of an algorithm to identify cardiovascular deaths from administrative health records: a multi-database population-based cohort study. OriginalPaper, BMC Health Serv. Res. 21 (1) (2021) 1–11, https://doi.org/10.1186/s12913-021-06762-0, 2021–07-31.
[25] L. TC, Derivation and validation of 10-year all-cause and cardiovascular disease mortality prediction model for middle-aged and elderly community-dwelling adults in Taiwan, PloS one 15 (9) (2020), https://doi.org/10.1371/journal.pone.0239063, 09/14/2020.
[26] S. SWA, A machine learning-based 1-year mortality prediction model after hospital discharge for clinical patients with acute coronary syndrome, Health Informatics J. 26 (2) (2020 Jun), https://doi.org/10.1177/1460458219871780, 2020.
[27] (CDC) CfDCaP, Data access - National Death Index, Updated 2021-02-08T01:30:29Z, https://www.cdc.gov/nchs/ndi/index.htm.
[28] R.B. Parikh, , Department of Medicine PSoM, Philadelphia University of Pennsylvania, Philadelphia Abramson Cancer Center UoP, Machine learning approaches to predict 6-month mortality among patients with cancer, JAMA Netw. Open 2 (10) (2021), https://doi.org/10.1001/jamanetworkopen.2019.15997.
[29] L. MS, Cause of death in clinical research: time for a reassessment? J. Am. Coll. Cardiol. 34 (3) (1999 Sep) https://doi.org/10.1016/s0735-1097(99)00250-8, 1999.