

RESEARCH ARTICLE

# External validation and comparison of two variants of the Elixhauser comorbidity measures for all-cause mortality

Yannick Fortin<sup>1,2\*</sup>, James A. G. Crispo<sup>1,2,3</sup>, Deborah Cohen<sup>2,4,5</sup>, Douglas S. McNair<sup>6</sup>, Donald R. Mattison<sup>1,7</sup>, Daniel Krewski<sup>1,2,7</sup>

**1** McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Ottawa, Ontario, Canada, **2** School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Ontario, Canada, **3** Fulbright Canada Student, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **4** Canadian Population Health Initiative (CPHI), Canadian Institute for Health Information (CIHI), Ottawa, Ontario, Canada, **5** Institute for Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada, **6** Cerner Corporation, Kansas City, Missouri, United States of America, **7** Risk Sciences International, Ottawa, Ontario, Canada

\* [yfort006@uottawa.ca](mailto:yfort006@uottawa.ca)



**OPEN ACCESS**

**Citation:** Fortin Y, Crispo JAG, Cohen D, McNair DS, Mattison DR, Krewski D (2017) External validation and comparison of two variants of the Elixhauser comorbidity measures for all-cause mortality. PLoS ONE 12(3): e0174379. <https://doi.org/10.1371/journal.pone.0174379>

**Editor:** Thanh G Phan, Monash University, AUSTRALIA

**Received:** July 19, 2016

**Accepted:** March 8, 2017

**Published:** March 28, 2017

**Copyright:** © 2017 Fortin et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Data used in this study are available from the Harvard Dataverse repository at: <https://dataverse.harvard.edu/privateurl.xhtml?token=dc1fca39-fa99-4b43-b5d1-1ae6ddcab7c1>.

**Funding:** This study was supported by doctoral training awards from the Fonds de recherche du Québec en santé (FF11 23D, <http://www.frqs.gouv.qc.ca>), the University of Ottawa ([www.uottawa.ca](http://www.uottawa.ca)), and the Ontario Ministry of Advanced Education and Skills Development (<https://osap.gov.on.ca>).

## Abstract

Assessing prevalent comorbidities is a common approach in health research for identifying clinical differences between individuals. The objective of this study was to validate and compare the predictive performance of two variants of the Elixhauser comorbidity measures (ECM) for in-hospital mortality at index and at 1-year in the Cerner Health Facts® (HF) U.S. database. We estimated the prevalence of select comorbidities for individuals 18 to 89 years of age who received care at Cerner contributing health facilities between 2002 and 2011 using the AHRQ (version 3.7) and the Quan Enhanced ICD-9-CM ECMs. External validation of the ECMs was assessed with measures of discrimination [c-statistics], calibration [Hosmer–Lemeshow goodness-of-fit test, Brier Score, calibration curves], added predictive ability [Net Reclassification Improvement], and overall model performance [R<sup>2</sup>]. Of 3,273,298 patients with a mean age of 43.9 years and a female composition of 53.8%, 1.0% died during their index encounter and 1.5% were deceased at 1-year. Calibration measures were equivalent between the two ECMs. Calibration performance was acceptable when predicting in-hospital mortality at index, although recalibration is recommended for predicting in-hospital mortality at 1 year. Discrimination was marginally better with the Quan ECM compared the AHRQ ECM when predicting in-hospital mortality at index ( $c_{\text{Quan}} = 0.887$ , 95% CI: 0.885–0.889 vs.  $c_{\text{AHRQ}} = 0.880$ , 95% CI: 0.878–0.882;  $p < .0001$ ) and at 1-year ( $c_{\text{Quan}} = 0.884$ , 95% CI: 0.883–0.886 vs.  $c_{\text{AHRQ}} = 0.880$ , 95% CI: 0.878–0.881,  $p < .0001$ ). Both the Quan and the AHRQ ECMs demonstrated excellent discrimination for in-hospital mortality of all-causes in Cerner Health Facts®, a HIPAA compliant observational research and privacy-protected data warehouse. While differences in discrimination performance between the ECMs were statistically significant, they are not likely clinically meaningful.

The McLaughlin Centre for Population Health Risk Assessment ([www.mclaughlincentre.ca](http://www.mclaughlincentre.ca)), the Cerner Corporation (<http://www.cerner.com>), and Risk Sciences International ([www.risksciences.com](http://www.risksciences.com)) supported the study indirectly in the form of salaries to the authors [YF, JAGC, DSM, DRM, DK]. The funders were not directly involved in the decision to prepare and submit this study for publication. The Cerner Corporation and Risk Sciences International are commercial entities.

**Competing interests:** Mr. Yannick Fortin, Dr. James A.G. Crispo, and Dr. Deborah Cohen report no competing interests. Dr. Douglas S. McNair is the President of Cerner Math Inc. and has ownership interest in the Cerner Corporation. Dr. Donald R. Mattison serves as Chief Medical Officer of Risk Sciences International, a company formed in partnership with the University of Ottawa in 2006 ([www.risksciences.com](http://www.risksciences.com)). RSI undertakes risk assessment work for public and private sector clients in Canada and abroad. RSI has not conducted prior work on the subject of the present research paper. Dr. Daniel Krewski is Chief Risk Scientist and CEO of Risk Sciences International. Since 2009, Dr. Krewski has held a Natural Sciences and Engineering Council of Canada (NSERC) Industrial Research Chair in Risk Science, through a peer-reviewed university-industry partnerships program administered by NSERC. None of the industrial partners in this study project were from the pharmaceutical industry. The authors' adherence to PLOS ONE policies are not altered by the commercial affiliation with the Cerner Corporation and Risk Sciences International.

## Introduction

With data on over 47 million unique patients who received care at nearly 500 US care facilities since 2000, the Cerner Health Facts® (HF) electronic health record database is a rich source of data available for epidemiological and health services research [1]. In addition to demographic and payer data, HF contains longitudinal diagnostic, procedure, pharmacy, and laboratory information on individuals receiving care within Cerner data networks.

To date, the predictive performance of commonly used comorbidity risk adjustment methods have yet to be corroborated in HF. Measures of comorbidity are useful tools for controlling for variation in overall patient health or adjusting for case-mix in epidemiological studies using electronic health data [2–5]. They are also used in observational drug effectiveness, health services, and outcomes studies when the unit of analysis cannot be appropriately randomized [6–10]. Measures of patient comorbidity have shown to be good predictors of short- and long-term mortality, hospital costs, length of stay (LOS), and readmission [11–14]. Failure to take patient comorbidity into account may lead to biased analyses, possibly due to confounding or systematic differences in health status among populations. Validated comorbidity measures may be used to address this issue.

In its simplest form, measures of comorbidity are aggregates of diagnostic codes used to identify the prevalence of predetermined health conditions in individuals documented in health data sources such as electronic health records. Of the many comorbidity measures validated for use with electronic health data, the original Elixhauser [15] comorbidity measures (ECM) are frequently reported as having greater predictive performance for short- and long-term mortality than competing models [11, 12, 16, 17]. The ECM target 30 medical, psychiatric, and lifestyle-related health conditions that are negatively associated with adverse health outcomes: congestive heart failure, cardiac arrhythmia, valvular disease, pulmonary circulation disorders, peripheral vascular disorders, hypertension (un/complicated), paralysis, neurological disorders, chronic pulmonary disease, uncomplicated diabetes, complicated diabetes, hypothyroidism, renal failure, liver disease, peptic ulcer disease without bleeding, aids/HIV, lymphoma, metastatic cancer, solid tumor without metastasis, rheumatoid arthritis/collagen vascular diseases, coagulopathy, obesity, weight loss, fluid and electrolyte disorders, blood loss anemia, deficiency anemia, alcohol abuse, drug abuse, psychoses, and depression. A few years after the publication of the original ECM, Quan et al. [18] and the Agency for Healthcare Research and Quality (AHRQ) [19] separately developed revised ECM variants, with Quan and colleagues reporting enhanced predictive performance for in-hospital mortality compared to both the original ECM and the AHRQ ECM (version 3.0). Studies have not examined whether the Quan variant is superior to the most recent AHRQ ECM (version 3.7) at predicting in-hospital mortality and in-hospital mortality at 1-year despite differences in the ICD-9-CM codes used to identify prevalent health conditions by each variant. The AHRQ ECM also differs from the Quan ECM by the exclusion of diagnostic codes making up a cardiac arrhythmia health condition group. Neither the Quan nor the AHRQ ECMs have been validated in HF.

The objectives of this study were to conduct an external validation and compare the performance of the Quan and AHRQ (version 3.7) ECMs for predicting in-hospital mortality of all-causes at index and at 1-year in HF.

## Methods

### Data source

Study data were derived from inpatient and emergency care encounters in Cerner Health Facts® (Kansas City, MO, USA), an administrative health database compliant with the US

Health Insurance Portability and Accountability Act (HIPAA). Approximately 500 health care facilities have contributed patient-level clinical encounter data to HF since January 2000. Data contributors range in size from those with fewer than five beds to those with over 500 and are located throughout the U.S. with a greater proportion situated in the Northeastern region of the United States [20]. University affiliated teaching hospitals comprise 40% of data contributors and they contribute more than 60% of all health encounters. Contributing health care facilities are categorized by teaching status, population density, bed size, and census region. HF data includes diagnoses recorded during emergency department (ED) visits, outpatient care, and hospitalizations, pharmacy orders, surgical procedures, laboratory and microbiology tests, and clinical procedures [1].

This study was approved by the Office for Research Ethics and Integrity at the University of Ottawa.

### Study population and index encounter

Individuals 18 to 89 years of age at the time of an ED or inpatient encounter at any HF contributing facility between January 2002 and December 2012 were eligible for inclusion. One ED or inpatient encounter was randomly selected between January 1<sup>st</sup>, 2002 and December 31<sup>st</sup>, 2011 as the index encounter for each individual. Outpatient visits were excluded as possible index encounters since deaths, the study outcome, are relatively infrequent during outpatient care [21, 22]. Persons younger than 18 years were excluded due to the relatively lower prevalence of Elixhauser health conditions and mortality in this population. Individuals 90 years or older are assigned to a single category in HF in order to comply with HIPAA requirements and were excluded to ensure age remained continuous in our analyses. Care recipients who transferred to or from any other health facility during the index encounter were excluded to avoid cases with higher potential for missing information [23]. Patient characteristics included sex, age at the index encounter, health insurance status, and race restricted to Caucasians, African Americans, Hispanics, and Asians. Health insurance status was categorized using AHRQ recommendations [24], including private, Medicaid, Medicare, uninsured [self-pay], other (TRICARE-CHAMPUS, international plan, research funded, Title V, worker's compensation), or missing. For reasons unknown, some HF contributing health care facilities voluntarily withhold information on the health insurance status of their patients, which leads to a significant proportion (>40%) of missing values.

### Elixhauser comorbidity measures and variants

The original ECM [15] comprises binary indicators for the diagnosis of 30 clinical conditions, each defined by a combination of codes according to the International Classification of Disease, Ninth Edition (ICD-9) [25]. Two variants of the original ECM were compared with the study sample; the AHRQ's latest comorbidity software (version 3.7) [19] and Quan et al.'s [18] Enhanced ICD-9-CM classification of comorbidities. The presence or absence of ECM comorbidities was assessed by examining the ICD-9-CM diagnostic codes recorded during the index encounter.

### Outcomes

The primary study outcomes were inhospital mortality of any causes during the index encounter and at 1-year. Inhospital mortality at 1-year was defined as a death recorded in a discharge abstract for an ED visit or an inpatient admission during the year that followed the admission date of the index encounter. Deaths recorded during the index encounter were therefore included in the mortality at 1-year outcome.

## Data analysis

The prevalence of comorbidities was described with counts and percentages while frequency differences between the ECMs were compared with McNemar's test [26]. To quantify the size and the clinical importance of the observed differences in the prevalence of comorbidities across ECMs, we derived Cohen's  $h$  [27]. Multiple logistic regression was used to predict the risk of mortality outcomes and output overall measures of model performance. External validation of the ECMs in HF was accomplished by deriving measures of calibration and discrimination for every predictive model, outcome, and sample combination considered [28, 29]. Model discrimination was assessed using the Area under the Receiver Operating Characteristic (ROC) curve (AUROC), an indicator of the ability of the ECMs to discriminate between the mortality statuses [30, 31]. The AUROC is often referred to as the concordance index number (c-statistic) and ranges between 0.5 [no discrimination] and 1.0 [perfect discrimination], with values above 0.7, 0.8, and 0.9 considered reasonable, strong, and exceptional, respectively [32]. The discrimination performance of each ECM was compared to 1) a baseline model, and 2) the competing ECM. Predictors in the baseline model were limited to sex and age at the index encounter to align with prior ECM validation and comparison studies [33, 34]. Differences in the AUROC between the fitted models were tested using the ROC and ROCCONTRAST statements in SAS [30]. The ROCCONTRAST option is an algorithm based on the non-parametric Mann-Whitney statistics developed by DeLong et al. [35] for comparing the significance of differences between correlated ROC curves.

Model calibration measures included the Hosmer–Lemeshow goodness-of-fit test, which evaluates the degree of agreement between the predicted and observed risk of in-hospital mortality [36]. The Hosmer–Lemeshow goodness-of-fit test outputs a Pearson chi-square score with a corresponding p-value: rejection of the null hypothesis indicates an imperfect correlation between predicted and observed values. Calibration plots displaying predicted in-hospital mortality probabilities on the x-axis and observed in-hospital mortality frequencies on the y-axis were generated to visually inspect calibration performance across risk deciles. The plots were enhanced using a smoothing spline function. Brier scores, which equate to the mean squared difference between predicted probabilities and observed outcomes, were included as measures of model accuracy [with lower Brier scores reflecting greater accuracy] [37]. Explained variation was reported in terms of the adjusted  $R^2$ .

It has been argued that the AUROC can be of limited value when comparing small incremental differences between predictive models [38]. To quantify the net improvement in predictive ability of the ECM that achieved the highest level of discrimination over the ECM with the lowest level of discrimination, we computed the net reclassification improvement (NRI) measure [39, 40]. Category-free NRI ( $NRI > 0$ ) assesses whether individuals are reclassified correctly in a prognosis model compared to a reference model.  $NRI > 0$  is a quantification of the net correct changes in model-based probabilities for both events [where improvement equates to increased probabilities of the outcome] and non-events [where improvement equates to decreased probabilities of the outcome] [41]. The NRI was implemented without risk categories because this approach allows for universal comparisons, it is robust against changing event-rates, and it was the most objective approach available in light of the insufficient evidence for meaningful risk categories for all-cause mortality in the literature [40].  $NRI > 0$  values were computed using a SAS macro developed by Kennedy et al. [42] and are reported with the percentage of events and non-events correctly reclassified. Statistical analyses were completed with SAS 9.4 (SAS Institute Inc., Cary, NC, USA).

## Risk groups

Recent history of hospitalization or emergency department use is associated with increased risk of hospital readmission and death [43, 44]. To further explore the utility of the ECMs, measures of discrimination and calibration performance were generated for both high and low risk patient groups. Individuals with evidence of one or more inpatient stay in the 12 months preceding the index encounter, or three or more emergency department visits in the 3 months preceding the index encounter, were defined as high risk. Patients that did not satisfy the high-risk criteria were assigned to the low risk group.

## Sensitivity analyses

Admissible index encounters in this study included ED visits and inpatient stays. It is reasonable to hypothesize that persons admitted for inpatient stays would generally be at greater risk of inhospital death than persons visiting the emergency department. To investigate potential differences in ECM performance by index encounter type, complimentary validation analyses were performed on index encounters recorded as ED visits and inpatient stays separately.

## Results

We identified 3,273,298 unique individuals who satisfied our inclusion criteria and received care at a HF care facility between 2002 and 2011. Mean age was 43.9 years and women were the majority (53.8%) (Table 1). Individuals were primarily Caucasians (72.3%), with others identified as African American (21.7%), Hispanic (4.5%), or Asians (1.5%). Index encounters were reported by health care institutions from the four US census regions: Northeast (36.1%), Midwest (19.8%), South (32.9%), and West (11.2%). Privately insured individuals comprised 48.3% of non-missing payer class cases. Most patients in the sample were classified as low risk (92.0%). Approximately two-thirds of index encounters were ED visits (67.4%). A total of 31,298 (1.0%) and 50,215 (1.5%) inhospital deaths of all-cause were recorded during the index encounter and at 1-year, respectively. As expected, high risk patients had a greater frequency of inhospital mortality than low risk patients at index [1.9% vs 0.9%,  $\chi^2$  (1, N = 3,273,298) = 2,792.1,  $p < .0001$ ], and at 1 year [4.2% vs 1.3%,  $\chi^2$  (1, N = 3,273,298) = 13,509.1,  $p < .0001$ ].

In descending order, hypertension, uncomplicated diabetes, chronic pulmonary disease, and fluid and electrolyte disorders were the most prevalent conditions identified by both ECMs (Table 2). Excluding HIV/AIDS, differences in prevalence between the ECMs for every health condition group were statistically significant based on McNemar's test,  $p < 0.0001$ . However, for most of the conditions assessed, the variation in prevalence between the ECMs differed by less than 1%, with the greatest differences observed for deficiency anemia (2.58%) and psychoses (1.22%). According to the interpretation criteria suggested by Cohen [27], only the deficiency anemia group demonstrated a small ( $h = 0.203$ ) practically meaningful difference in prevalence by ECM.

Table 3 reports the performance measures of discrimination and calibration for the study sample and by patient risk groups. When predicting inhospital mortality during the index encounter, the Quan model ( $c = 0.887$ , 95% CI: 0.885–0.889) had negligible but significantly higher discrimination than the AHRQ ( $c = 0.880$ , 95% CI: 0.878–0.882) and baseline ( $c = 0.820$ , 95% CI: 0.818–0.822) models,  $p < .0001$ . Similar results were obtained for predicting inhospital mortality at 1-year, the discrimination performance of the Quan model ( $c = 0.884$ , 95% CI: 0.883–0.886) marginally exceeded the discrimination performance of the AHRQ ( $c = 0.880$ , 95% CI: 0.878–0.881) and baseline ( $c = 0.826$ , 95% CI: 0.824–0.827) models,  $p < .0001$ . Model discrimination for the mortality outcomes remained strong (AUROC > 0.8) and the observed advantage of the Quan ECM over the AHRQ ECM was confirmed in both risk group samples.



**Table 1. Patient demographic and index encounter characteristics, N = 3,273,298.**

Characteristic	N (%)
<b>Sex</b>	
Female	1,761,525 (53.8)
Male	1,511,773 (46.2)
<b>Age (Years)</b>	
Mean ± SE	43.9 ± 0.01
<b>Race</b>	
Caucasian	2,366,665 (72.3)
African American	711,051 (21.7)
Hispanic	146,877 (4.5)
Asian	48,705 (1.5)
<b>Insurance Status</b>	
Private	795,449 (24.3)
Medicare	370,701 (11.3)
Medicaid	248,009 (7.6)
Uninsured	378,536 (11.6)
Other	139,745 (4.3)
Missing	1,340,858 (41.0)
<b>Risk Group</b>	
Low	3,010,916 (92.0)
High	262,382 (8.0)
<b>Inhospital Mortality</b>	
Deaths	31,298 (1.0)
<b>Inhospital Mortality at 1-Year</b>	
Deaths	50,215 (1.5)
<b>Care Setting, Index Encounter</b>	
ED Visit	2,204,680 (67.4)
Inpatient Stay	1,068,618 (32.6)
<b>Census Region</b>	
Northeast	1,180,270 (36.1)
Midwest	648,644 (19.8)
South	1,077,965 (32.9)
West	366,419 (11.2)

**Abbreviations:** ED, emergency department; SE, Standard error.

<https://doi.org/10.1371/journal.pone.0174379.t001>

ROC plots displaying the minor discrimination advantage of the Quan ECM over the AHRQ ECM are provided in Figs 1 and 2.

There were no differences in Brier scores between competing ECMs, irrespective of the predicted outcome. The Brier scores were consistently lower when predicting in-hospital mortality at index than at 1 year. These findings might indicate that the ECMs have better calibration when predicting the former outcome than when predicting the latter. The calibration plots reported in Figs 3 and 4 show good agreement between predicted and observed risk of in-hospital mortality at index. However, the level agreement between the predicted and observed risk of in-hospital mortality at 1 year were less satisfactory, suggesting the need for recalibration. As the observed risk of mortality at 1 year increased, the ECMs increasingly over-predicted the outcome. Results from the Hosmer–Lemeshow goodness-of-fit test indicated imperfect agreement between expected and observed risk, irrespective of the ECM–outcome combination

**Table 2. Prevalence of comorbid conditions by ECM variant, N = 3,273,298.**

Condition	Quan, N (%)	AHRQ, N (%)	McNamar's Test P Value	Cohen's h
Hypertension	572,139 (17.48)	573,457 (17.52)	< .0001	0.001
Chronic Pulmonary Disease	256,170 (7.83)	248,676 (7.60)	< .0001	0.009
Diabetes Uncomplicated	226,807 (6.93)	227,815 (6.96)	< .0001	0.001
Fluid and Electrolyte Disorders	187,321 (5.72)	186,282 (5.69)	< .0001	0.001
Cardiac Arrhythmia	174,656 (5.34)	na	na	na
Depression	113,659 (3.47)	89,410 (2.73)	< .0001	0.043
Congestive Heart Failure	99,280 (3.03)	90,775 (2.77)	< .0001	0.015
Alcohol Abuse	89,460 (2.73)	86,307 (2.64)	< .0001	0.006
Hypothyroidism	85,493 (2.61)	83,926 (2.56)	< .0001	0.003
Obesity	79,680 (2.43)	80,996 (2.47)	< .0001	0.003
Other Neurological Disorders	76,237 (2.33)	97,360 (2.97)	< .0001	0.040
Drug Abuse	61,260 (1.87)	59,029 (1.80)	< .0001	0.005
Renal Failure	60,238 (1.84)	57,231 (1.75)	< .0001	0.007
Solid Tumor without Metastasis	56,376 (1.72)	56,519 (1.73)	< .0001	0.000
Valvular Disease	51,250 (1.57)	21,907 (0.67)	< .0001	0.087
Peripheral Vascular Disorders	41,651 (1.27)	43,410 (1.33)	< .0001	0.005
Diabetes Complicated	34,016 (1.04)	34,111 (1.04)	< .0001	0.000
Liver Disease	31,307 (0.96)	21,910 (0.67)	< .0001	0.032
Psychoses	29,656 (0.91)	69,664 (2.13)	< .0001	0.102
Coagulopathy	27,019 (0.83)	27,330 (0.83)	< .0001	0.001
Rheumatoid Arthritis/collagen	24,143 (0.74)	22,407 (0.68)	< .0001	0.006
Metastatic Cancer	22,836 (0.70)	23,004 (0.70)	< .0001	0.001
Weight Loss	21,219 (0.65)	19,429 (0.59)	< .0001	0.007
Pulmonary Circulation Disorders	21,084 (0.64)	20,894 (0.64)	< .0001	0.001
Deficiency Anemia	19,745 (0.60)	104,246 (3.18)	< .0001	0.203*
Paralysis	13,707 (0.42)	19,001 (0.58)	< .0001	0.023
Blood Loss Anemia	10,884 (0.33)	24,487 (0.75)	< .0001	0.058
Peptic Ulcer Disease excl. bleeding	8,251 (0.25)	507 (0.02)	< .0001	0.076
Lymphoma	6,778 (0.21)	6,920 (0.21)	< .0001	0.001
AIDS/HIV	4,246 (0.13)	4,246 (0.13)	1.000	0.000

\* Described as a small effect size according to the interpretation criteria suggested by Cohen (1988).

<https://doi.org/10.1371/journal.pone.0174379.t002>

assessed. This is expected given the large study sample and previously reported simulation results from Kramer et al. [36] showing the Hosmer-Lemeshow test to be particularly sensitive to sample size: even with a minor deviation (0.4%) from perfect fit between expected and observed risk, studies with sample sizes of 50,000 or more observations rejected the null hypothesis 100% of the time.

Explained variation ( $R^2$ ) was consistently higher, by approximately 1 to 2% with the Quan ECM compared to the AHRQ ECM across patient groups and mortality outcomes. Measures of discrimination (AUROC), calibration (Brier scores), and overall performance ( $R^2$ ) were consistently better in the low risk patient group compared to the high risk patient group.

Net reclassification improvements were observed by the Quan ECM over the AHRQ ECM for the full sample, in high and low risk patients, and in patients with index encounter limited to inpatient stays. However, the magnitude of these improvements was low to moderate, between 0.35 to 0.62, on a possible NRI range of -2 to 2. The positive NRI scores observed, and

**Table 3. Measures of discrimination and calibration performance by ECM and mortality outcome.**

		External Validation			
		Inhospital Mortality at Index		Inhospital Mortality at 1 Year	
		Quan	AHRQ	Quan	AHRQ
<b>All Patients N = 3,273,298</b>	AUROC <sup>a</sup> (95% CI)	0.887 (0.885,0.889) <sup>G</sup>	0.880 (0.878,0.882) <sup>G</sup>	0.884 (0.883,0.886) <sup>G</sup>	0.880 (0.878,0.881) <sup>G</sup>
	HL Test <sup>b</sup>	485.5*	459.8*	890.1*	879.1*
	Brier Score <sup>c</sup>	0.009	0.009	0.014	0.014
	R <sup>2</sup> <sup>d</sup>	24.9	23.1	25.1	24.0
	NRI>0 <sup>e</sup> (95% CI)	0.6115 (0.6006,0.6224)*		0.5234 (0.5149,0.5318)*	
	Reclassification, E—NE <sup>f</sup>	-18%–79%		-27%–80%	
	Deaths (%)	31,298 (1.0)		50,215 (1.5)	
<b>High Risk Patients<sup>h</sup> N = 262,382</b>	AUROC <sup>a</sup> (95% CI)	0.870 (0.865,0.874) <sup>G</sup>	0.861 (0.857,0.866) <sup>G</sup>	0.834 (0.830,0.837) <sup>G</sup>	0.830 (0.826,0.833) <sup>G</sup>
	HL Test <sup>b</sup>	106.6*	96.7*	271.5*	261.1*
	Brier Score <sup>c</sup>	0.018	0.018	0.037	0.037
	R <sup>2</sup> <sup>d</sup>	22.7	21.4	20.7	20.0
	NRI>0 <sup>e</sup> (95% CI)	0.4725 (0.4450,0.5000)*		0.3566 (0.3384,0.3747)*	
	Reclassification, E—NE <sup>f</sup>	-15%–62%		-28%–64%	
	Deaths (%)	5,035 (1.9)		11,043 (4.2)	
<b>Low Risk Patients<sup>h</sup> N = 3,010,916</b>	AUROC <sup>a</sup> (95% CI)	0.886 (0.884,0.889) <sup>G</sup>	0.879 (0.877,0.881) <sup>G</sup>	0.885 (0.883,0.886) <sup>G</sup>	0.880 (0.878,0.882) <sup>G</sup>
	HL Test <sup>b</sup>	378.0*	410.6*	619.4*	616.2*
	Brier Score <sup>c</sup>	0.008	0.008	0.012	0.012
	R <sup>2</sup> <sup>d</sup>	24.8	22.9	25.0	23.8
	NRI>0 <sup>e</sup> (95% CI)	0.6183 (0.6064,0.6302)*		0.5384 (0.5288,0.548)*	
	Reclassification, E—NE <sup>f</sup>	-18%–80%		-27%–80%	
	Deaths (%)	26,263 (0.9)		39,172 (1.3)	

\* P-value < 0.001. E = Events. NE = Non-events.

<sup>a</sup> Area under the Receiver Operating Characteristic (ROC) curve (AUROC). AUROC is a measure of discrimination ranging from 0.5 (zero discrimination) to 1.0 (perfect discrimination).

<sup>b</sup> Pearson chi-square value derived from the Hosmer–Lemeshow goodness-of-fit test [32].

<sup>c</sup> Measure of predictive accuracy, greater accuracy is reflected by lower score.

<sup>d</sup> R-squared, explained variation, displayed in percentage.

<sup>e</sup> Category-free net reclassification improvement using the AHRQ ECM as the reference model.

<sup>f</sup> E–NE, percentage of events (E) and non-events (NE) correctly reclassified by the Quan ECM compared to the AHRQ ECM.

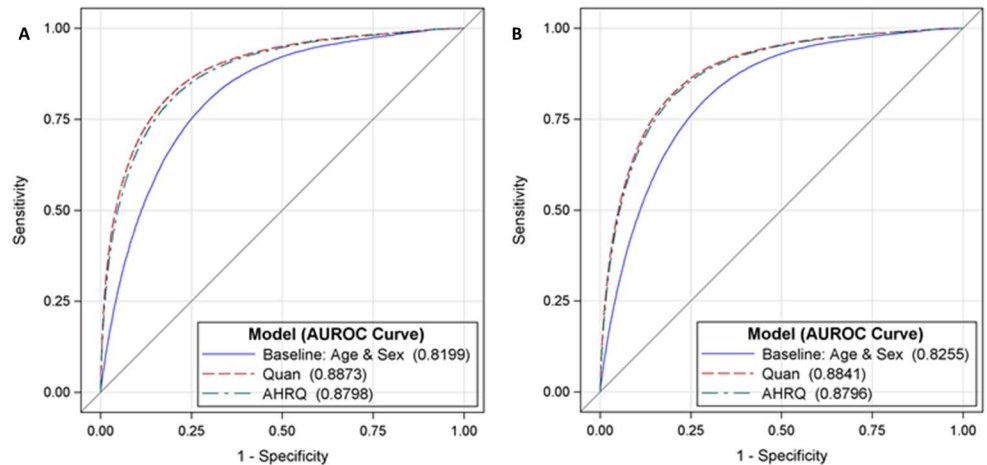
<sup>g</sup> AUROC curve differed significantly from the baseline model limited to age and sex ( $p < 0.0001$ ), and from the competing ECM ( $p < 0.0001$ ). Differences between AUROC curves were evaluated with the Mann-Whitney U test approach developed by DeLong et al. [35]. In the unstratified sample, the baseline model had an AUROC of 0.820 (95% CI 0.818–0.822) for in-hospital mortality at index, and 0.826 (95% CI 0.824–0.827) for in-hospital mortality at 1 year. For high risk patients, the baseline model had an AUROC of 0.770 (95% CI 0.764–0.775) for in-hospital mortality at index, and 0.755 (95% CI 0.751–0.759) for in-hospital mortality at 1 year. For low risk patients, the baseline model had an AUROC of 0.821 (95% CI 0.819–0.823) for in-hospital mortality at index, and 0.828 (95% CI 0.826–0.830) for in-hospital mortality at 1 year.

<sup>h</sup> High risk patients had one or more inpatient stay in the 12 months preceding the index encounter or three or more emergency department visits in the 3 months preceding the index encounter. Patients that did not satisfy the high-risk criteria were assigned to the low risk group.

<https://doi.org/10.1371/journal.pone.0174379.t003>

the apparent greater predictive performance of the Quan ECM, were driven principally by improvements in model specificity (the correct reclassification of non-events). In every sample-outcome combination examined, the percentage of events correctly reclassified by the





**Fig 1. AUROC comparisons by ECM for predicting in-hospital mortality at index [A] and at 1 year [B].** ROC = receiver operating characteristic, AUROC = area under the receiver operating characteristic.

<https://doi.org/10.1371/journal.pone.0174379.g001>

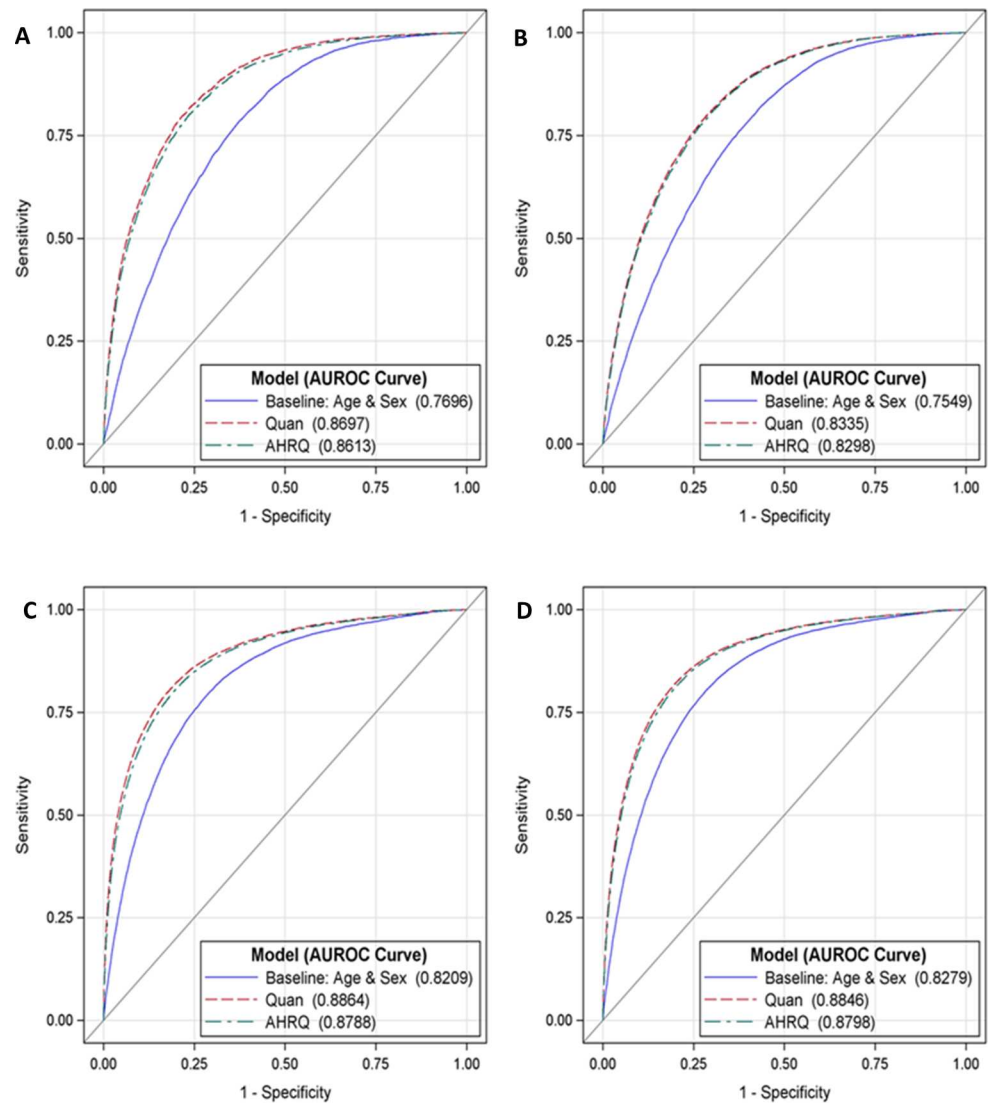
Quan ECM was negative, possibly indicating reduced sensitivity compared to the AHRQ ECM. In the sample limited to persons whose index encounter was an emergency department visit, this reduced sensitivity combined with minimal improvement in specificity by the Quan ECM compared to the AHRQ ECM resulted in negative NRI.

### Sensitivity analyses

Results from the sensitivity analyses are available as supplementary material (S1 Table, S1 and S2 Figs). As hypothesized, patients whose index encounter was an inpatient stay had a greater risk of in-hospital mortality at index [2.4% vs 0.3%,  $\chi^2(1, N = 3,273,298) = 33,138.7, p < .0001$ ], and at 1 year [3.6% vs 0.6%,  $\chi^2(1, N = 3,273,298) = 43,585.6, p < .0001$ ] than patients whose index encounter was an ED visit. Performance measures in the analyses stratified by index encounter type mimicked the trends reported for the unstratified sample, except for explained variation which was 48 to 69 percent lower in the ED visit sample than in the inpatient sample. Measures of discrimination (AUROC) and calibration (Brier scores) were superior in the inpatient sample compared to the ED visit sample.

### Discussion

We conducted an external validation and compared the ability of the Quan and AHRQ ECMs to predict in-hospital mortality at index and at 1-year in the Cerner HF database. In a prior study, the Quan ECM demonstrated superior predictive performance over the AHRQ version 3.0 ECM for in-hospital mortality at index in a Canadian population with universal health coverage [18]. The current study expands on prior findings and demonstrates the performance advantage of the Quan ECM over the AHRQ version 3.7 ECM for in-hospital mortality at index and at 1-year in Cerner Health Facts®. This is the first study to evaluate any diagnostic-based risk adjustment methods in HF and to confirm the excellent discrimination performance of both the Quan and AHRQ ECMs in a multi-payer US health data source. While significant, increased discrimination performance of the Quan ECM over the AHRQ ECM did not exceed 1% for any of the mortality outcomes after the inclusion of baseline variables age and sex. It is therefore unlikely that the observed differences between the ECMs are clinically meaningful. The marginal performance improvement in discrimination and explained variance observed between the ECMs may be a consequence of the large sample available for

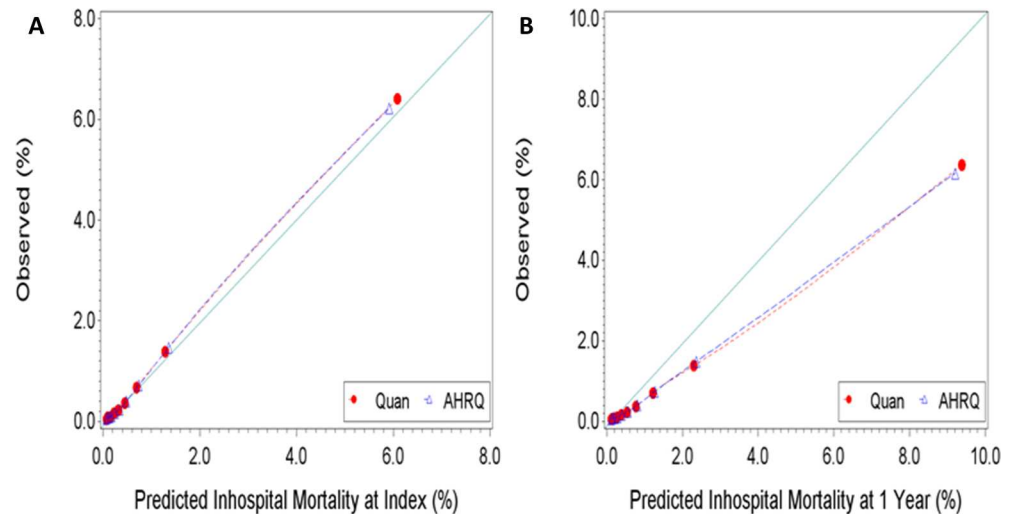


**Fig 2. AUROC comparisons by ECM for predicting in-hospital mortality at index [A] and at 1 year [B] in high risk patients, and in-hospital mortality at index [C] and at 1 year [D] in low risk patients.** High risk patients had one or more inpatient stay in the 12 months preceding the index encounter or three or more emergency department visits in the 3 months preceding the index encounter. Patients that did not satisfy the high-risk criteria were assigned to the low risk group. ROC = receiver operating characteristic, AUROC = area under the receiver operating characteristic.

<https://doi.org/10.1371/journal.pone.0174379.g002>

analysis and might not be reproducible in smaller HF subsets or patient subpopulations. In this study, evidence of superior predictive performance by the Quan ECM was demonstrated in an undifferentiated patient population, in patient groups stratified by risk of hospital readmission and death, and in patients stratified by their index encounter type [ED visits and inpatient stays].

Visual inspection of the calibration plots for the Quan and AHRQ ECMs revealed noticeable levels of disagreement between predicted and observed risk of in-hospital mortality at 1 year. Lower calibration performance appeared more pronounced in the high risk patient group compared to the low risk patient group. To improve accuracy, we recommend that the ECMs be recalibrated specifically for predictions of in-hospital mortality at 1 year in HF.



**Fig 3. Observed versus predicted risk of in-hospital mortality [A] at index and [B] at 1 Year.** Perfect calibration is represented by the full line with a slope of 1 starting at the origin.

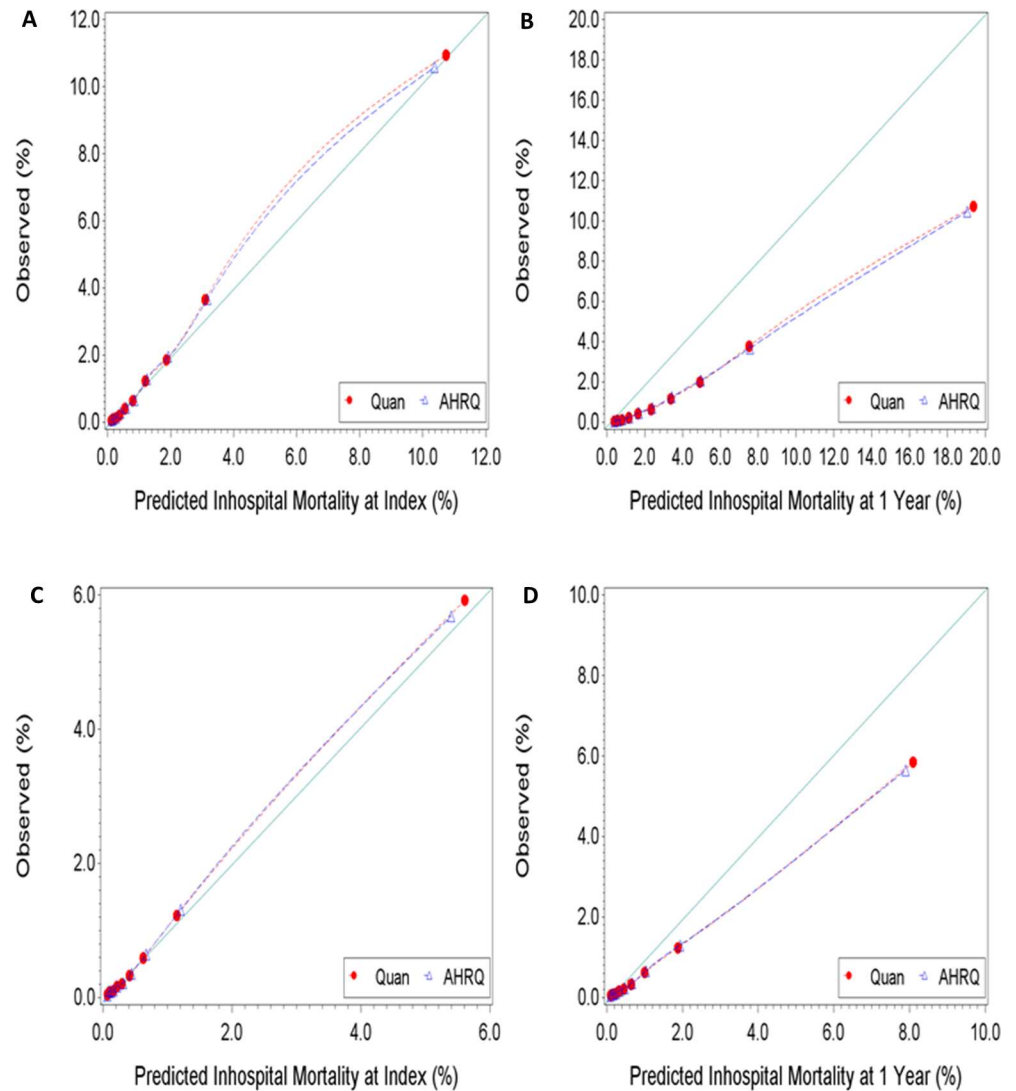
<https://doi.org/10.1371/journal.pone.0174379.g003>

The observed over-prediction of in-hospital mortality at 1 year by the ECMs likely results from a combination of factors ranging from suboptimal parameter selection to outcome misclassification.

The ICD-9 codes used to assess the prevalence of comorbid conditions by the competing ECMs resulted in minimal variations in disease frequencies; prevalence differed by less than 1% for the majority of conditions. The exclusion of cardiac arrhythmia from the AHRQ ECM may be responsible for the observed predictive performance differences. In the Quan ECM, cardiac arrhythmia was the fifth most prevalent condition and was significantly associated with increased odds of in-hospital mortality at index and at 1-year in the adjusted models (S2 Table).

This study has limitations. Like other EHR-derived data warehouse used for observational research that comply with the U.S. HIPAA law and regulations, HF adheres to de-identification procedures that prevent further linkage to registries such as the National Death Index and other health organizations outside the same HIPAA-covered entity. Since deaths were limited to those recorded during inpatient care and within HIPAA-covered networks, some misclassification of mortality at 1-year is to be expected from the deaths that occurred outside these settings. The patient de-identification process also implies that the diagnoses and combination of ICD codes used for estimating ECM prevalence could not be validated using chart re-abstracting methodology leaving doubts about their sensitivity and specificity in HF. Limiting the assessment of morbidities to a single index encounter, as opposed to including a look back period in the assessment of a patient's health, likely resulted in the misclassification of previously diagnosed health conditions as absent. Including a look back periods of one to two years generally improves the detection of prevalent health conditions [45, 46]. One explanation for this improvement is that look back periods limit bias in discharge abstract coding whereby secondary health conditions tend to be under recorded in patients treated for severe acute conditions and vice versa [15, 47]. A look back period was not included in this study to increase the comparability of results with the original Quan [18] paper and because our research group is currently conducting a parallel study to test the consequences of varying look back periods in HF.

HF was not primarily designed for research purposes [1]. Study findings are therefore subject to the same risks, biases, and limitations typically associated with research based on



**Fig 4. Observed versus predicted risk of in-hospital mortality at index [A] and at 1 year [B] for high risk patients, and in-hospital mortality at index [C] and at 1 year [D] for low risk patients.** Perfect calibration is represented by the full line with a slope of 1 starting at the origin. High risk patients had one or more inpatient stay in the 12 months preceding the index encounter or three or more emergency department visits in the 3 months preceding the index encounter. Patients that did not satisfy the high-risk criteria were assigned to the low risk group.

<https://doi.org/10.1371/journal.pone.0174379.g004>

electronic health data [48, 49]. These include potential for selection bias, missing or incomplete documentation, coding errors, misclassifications of diagnostic codes, record linkage errors due to interoperability issues, and duplication. Finally, recorded comorbid conditions could not be separated from conditions resulting from complications in care. Thus, it was impossible to evaluate the effects of excluding complications of care from our models.

The Quan and the AHRQ (version 3.7) ECMs were found to be practically equivalent in discriminating between short- and long-term in-hospital mortality outcomes in HF. While ECM calibration measures were satisfactory for predicting in-hospital mortality at index, recalibration of the ECMs is recommended to improve the predictive accuracy for in-hospital mortality at 1 year. These diagnostic-based risk adjustment tools should enhance

capacity for conducting quality observational studies and health services research using Health Facts® data.

## Supporting information

**S1 Fig. AUROC comparison by ECM for predicting inhospital mortality at index [A] and at 1 Year [B] for index encounters limited to emergency department visits, and inhospital mortality at index [C] and at 1 Year [D] for index encounters limited to inpatient stays.**

AUROC = area under the receiver operating characteristic, ROC = receiver operating characteristic.

(TIF)

**S2 Fig. Observed versus predicted risk of inhospital mortality at index [A] and at 1 Year [B] for index encounters limited to emergency department visits, and inhospital mortality at index [C] and at 1 year [D] for index encounters limited to inpatient stays.** Perfect calibration is represented by the full line with a slope of 1 starting at the origin.

(TIF)

**S1 Table. Measures of discrimination and calibration performance for inhospital mortality by index encounter type, ED visits and inpatient stays.** \* P-value < 0.001. ED = Index encounter is an emergency department visit. IS = Index encounter is an inpatient stay. E =

Events. NE = Non-events. <sup>a</sup> Area under the Receiver Operating Characteristic (ROC) curve (AUROC). AUROC is a measure of discrimination ranging from 0.5 (zero discrimination) to 1.0 (perfect discrimination). <sup>b</sup> Pearson chi-square value derived from the Hosmer–Lemeshow goodness-of-fit test [32]. <sup>c</sup> Measure of predictive accuracy, greater accuracy is reflected by lower score. <sup>d</sup> Generalized R-squared, explained variation, displayed in percentage. <sup>e</sup> Category-free net reclassification improvement with the AHRQ ECM as the reference model. <sup>f</sup> E-NE, percentage of events (E) and non-events (NE) correctly reclassified by the Quan ECM compared to the AHRQ ECM. <sup>g</sup> AUROC curve differed significantly from the baseline model limited to age and sex ( $p < 0.0001$ ), and from the competing ECM ( $p < 0.0001$ ). Differences between AUROC curves were evaluated with the Mann-Whitney U test approach developed by DeLong et al. (1988). For ED encounters, the baseline model had an AUROC of 0.804 (95% CI 0.799–0.810) for inhospital mortality at index, and 0.826 (95% CI 0.822–0.829) for inhospital mortality at 1 year. For IS encounters, the baseline model had an AUROC of 0.752 (95% CI 0.749–0.754) for inhospital mortality at index, and 0.754 (95% CI 0.752–0.756) for inhospital mortality at 1 year.

(DOCX)

**S2 Table. Adjusted odds ratios (95% CI) of Elixhauser conditions for inpatient mortality at index and at 1-year by ECM variant, N = 3,273,298.** Odds ratios are adjusted for baseline variables sex and age. The odds ratios reported are those that reached statistical significance ( $p < 0.05$ ). Abbreviation: CI, confidence intervals.

(DOCX)

## Acknowledgments

We wish to acknowledge the generous support of the Cerner staff for helping us appreciate the riches and complexities of the Health Facts® database. Preliminary findings for this study were presented in poster form at the 20th ISPOR Annual International Meeting in Philadelphia in May 2015. This paper reflects the opinions of the authors and not necessarily those of

the Canadian Institute for Health Information, the Cerner Corporation, or Risk Sciences International.

## Author Contributions

**Conceptualization:** YF JAGC DRM DK.

**Data curation:** YF JAGC.

**Formal analysis:** YF JAGC.

**Funding acquisition:** YF DK.

**Investigation:** YF DSM.

**Methodology:** YF JAGC DC DSM DRM DK.

**Project administration:** YF.

**Resources:** DSM DK.

**Software:** YF JAGC.

**Supervision:** DRM DK.

**Validation:** YF.

**Visualization:** YF.

**Writing – original draft:** YF JAGC DK.

**Writing – review & editing:** YF JAGC DC DSM DRM DK.

## References

1. B.R.I.D.G.E. to Data. Cerner Health Facts® Database (USA) Arlington: VA2014 [Available from: <http://www.bridgetodata.org/node/1789>.
2. van Gestel YR, Lemmens VE, Lingsma HF, de Hingh IH, Rutten HJ, Coebergh JWW. The hospital standardized mortality ratio fallacy: a narrative review. *Med Care*. 2012; 50(8):662–7. <https://doi.org/10.1097/MLR.0b013e31824ebd9f> PMID: 22410410
3. Schneeweiss S, Maclure M. Use of comorbidity scores for control of confounding in studies using administrative databases. *Int J Epidemiol*. 2000; 29(5):891–8. PMID: 11034974
4. Schneeweiss S, Seeger JD, Maclure M, Wang PS, Avorn J, Glynn RJ. Performance of Comorbidity Scores to Control for Confounding in Epidemiologic Studies using Claims Data. *Am J Epidemiol*. 2001; 154(9):854–64. PMID: 11682368
5. Fischer C, Lingsma H, van Leersum N, Tollenaar R, Wouters M, Steyerberg E. Comparing colon cancer outcomes: The impact of low hospital case volume and case-mix adjustment. *European Journal of Surgical Oncology (EJSO)*. 2015.
6. Bosetti C, Franchi M, Nicotra F, Ascitto R, Merlino L, La Vecchia C, et al. Insulin and other antidiabetic drugs and hepatocellular carcinoma risk: a nested case-control study based on Italian healthcare utilization databases. *Pharmacoepidemiol Drug Saf*. 2015.
7. Holmes HM, Luo R, Kuo Y-F, Baillargeon J, Goodwin JS. Association of potentially inappropriate medication use with patient and prescriber characteristics in Medicare Part D. *Pharmacoepidemiol Drug Saf*. 2013; 22(7):728–34. <https://doi.org/10.1002/pds.3431> PMID: 23494811
8. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiol Drug Saf*. 2012; 21:138–47. <https://doi.org/10.1002/pds.3231> PMID: 22552989
9. Suh HS, Hay JW, Johnson KA, Doctor JN. Comparative effectiveness of statin plus fibrate combination therapy and statin monotherapy in patients with type 2 diabetes: use of propensity-score and instrumental variable methods to adjust for treatment-selection bias. *Pharmacoepidemiol Drug Saf*. 2012; 21(5):470–84. <https://doi.org/10.1002/pds.3261> PMID: 22461130



10. French DD, Campbell R, Spehar A, Angaran DM. Benzodiazepines and injury: a risk adjusted model. *Pharmacoepidemiol Drug Saf.* 2005; 14(1):17–24. <https://doi.org/10.1002/pds.967> PMID: 15386711
11. Sharabiani MT, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. *Med Care.* 2012; 50(12):1109–18. <https://doi.org/10.1097/MLR.0b013e31825f64d0> PMID: 22929993
12. Chu Y-T, Ng Y-Y, Wu S-C. Comparison of different comorbidity measures for use with administrative data in predicting short- and long-term mortality. *BMC Health Serv Res.* 2010; 10(1):140.
13. Hall SF. A user's guide to selecting a comorbidity index for clinical research. *J Clin Epidemiol.* 2006; 59(8):849–55. <https://doi.org/10.1016/j.jclinepi.2005.11.013> PMID: 16828679
14. Yurkovich M, Avina-Zubieta JA, Thomas J, Gorenchtein M, Lacaille D. A systematic review identifies valid comorbidity indices derived from administrative health data. *J Clin Epidemiol.* 2015; 68(1):3–14. <https://doi.org/10.1016/j.jclinepi.2014.09.010> PMID: 25441702
15. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care.* 1998; 36(1):8–27. PMID: 9431328
16. Lix L, Quail J, Teare G, Acan B. Performance of comorbidity measures for predicting outcomes in population-based osteoporosis cohorts. *Osteoporos Int.* 2011; 22(10):2633–43. <https://doi.org/10.1007/s00198-010-1516-7> PMID: 21305268
17. Stukenborg GJ, Wagner DP, Connors AF Jr. Comparison of the performance of two comorbidity measures, with and without information from prior hospitalizations. *Med Care.* 2001; 39(7):727–39. PMID: 11458137
18. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi J-C, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care.* 2005; 43(11):1360–9. PMID: 16224307
19. AHRQ. Comorbidity Software, Version 3.7 Rockville, MD: Agency for Healthcare Research and Quality—Healthcare Cost and Utilization Project (HCUP); 2014 [Available from: <http://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp>].
20. Taylor R. Training on Cerner's Health Facts® Data Warehouse. In: Risk Sciences International Pharmacovigilance Group, editor. Ottawa, ON2014. p. 48.
21. Raymond EG, Grossman D, Weaver MA, Toti S, Winikoff B. Mortality of induced abortion, other outpatient surgical procedures and common activities in the United States. *Contraception.* 2014; 90(5):476–9. <https://doi.org/10.1016/j.contraception.2014.07.012> PMID: 25152259
22. Keyes GR, Singer R, Iverson RE, McGuire M, Yates J, Gold A, et al. Mortality in outpatient surgery. *Plast Reconstr Surg.* 2008; 122(1):245–50. <https://doi.org/10.1097/PRS.0b013e31817747fd> PMID: 18594412
23. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med Care.* 2009; 47(6):626–33. <https://doi.org/10.1097/MLR.0b013e31819432e5> PMID: 19433995
24. Barrett M, Lopez-Gonzalez L, Hines A, Andrews R, Jiang J. An Examination of Expected Payer Coding in HCUP Databases 2014. Available from: <http://www.hcup-us.ahrq.gov/reports/methods/methods.jsp>.
25. CDC. Classification of Diseases, Functioning, and Disability: International Classification of Diseases, Ninth Revision (ICD-9). In: Centers for Disease Control and Prevention, editor. Atlanta: GA2009.
26. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika.* 1947; 12(2):153–7. PMID: 20254758
27. Cohen J. *Statistical power analysis for the behaviour sciences.* 2 ed. Hillsdale, NJ: Laurence Earlbaum Associates; 1988.
28. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 2014; 14(1):40.
29. Moons KM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. *Ann Intern Med.* 2015; 162(1):W1–W73. <https://doi.org/10.7326/M14-0698> PMID: 25560730
30. Gonen M. Receiver operating characteristic (ROC) curves. *SAS Users Group International (SUGI).* 2006; 31:210–31.
31. Iezzoni LI. *Risk adjustment for measuring healthcare outcomes.* 2 ed. Chicago Ill: Health Administration Press; 1997.
32. Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied logistic regression.* 3 ed. Hoboken, NJ: John Wiley & Sons; 2013.
33. Zhu H, Hill MD. Stroke The Elixhauser Index for comorbidity adjustment of in-hospital case fatality. *Neurology.* 2008; 71(4):283–7. <https://doi.org/10.1212/01.wnl.0000318278.41347.94> PMID: 18645167

34. Li P, Kim MM, Doshi JA. Comparison of the performance of the CMS Hierarchical Condition Category (CMS-HCC) risk adjuster with the Charlson and Elixhauser comorbidity measures in predicting mortality. *BMC Health Serv Res.* 2010; 10(1):245.
35. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988:837–45. PMID: [3203132](#)
36. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med.* 2007; 35(9):2052–6. <https://doi.org/10.1097/01.CCM.0000275267.64078.B0> PMID: [17568333](#)
37. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly weather review.* 1950; 78(1):1–3.
38. Cook NR. Assessing the incremental role of novel and emerging risk factors. *Curr Cardiovasc Risk Rep.* 2010; 4(2):112–9. <https://doi.org/10.1007/s12170-010-0084-x> PMID: [20640227](#)
39. Pencina MJ, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med.* 2008; 27(2):157–72. <https://doi.org/10.1002/sim.2929> PMID: [17569110](#)
40. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med.* 2011; 30(1):11–21. <https://doi.org/10.1002/sim.4085> PMID: [21204120](#)
41. Pencina MJ, D'agostino RB, Vasan RS. Statistical methods for assessment of added usefulness of new biomarkers. *Clin Chem Lab Med.* 2010; 48(12):1703–11. <https://doi.org/10.1515/CCLM.2010.340> PMID: [20716010](#)
42. Kennedy K, Pencina M, editors. A SAS® macro to compute added predictive ability of new markers predicting a dichotomous outcome. *SouthEast SAS Users Group Annual Meeting Proceedings*; 2010.
43. van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Can Med Assoc J.* 2010; 182(6):551–7.
44. Lanière I, Couturier P, Dramé M, Gavazzi G, Lehman S, Jolly D, et al. Incidence and main factors associated with early unplanned hospital readmission among French medical inpatients aged 75 and over admitted through emergency units. *Age Ageing.* 2008; 37(4):416–22. <https://doi.org/10.1093/ageing/afn093> PMID: [18487268](#)
45. Preen DB, Holman CAJ, Spilsbury K, Semmens JB, Brameld KJ. Length of comorbidity lookback period affected regression model performance of administrative health data. *J Clin Epidemiol.* 2006; 59(9):940–6. <https://doi.org/10.1016/j.jclinepi.2005.12.013> PMID: [16895817](#)
46. Zhang JX, Iwashyna TJ, Christakis NA. The performance of different lookback periods and sources of information for Charlson comorbidity adjustment in Medicare claims. *Med Care.* 1999; 37(11):1128–39. PMID: [10549615](#)
47. Hughes JS, Iezzoni LI, Daley J, Greenberg L. How severity measures rate hospitalized patients. *J Gen Intern Med.* 1996; 11(5):303–11. PMID: [8725979](#)
48. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. *J Clin Epidemiol.* 2011; 65(2):126–31. <https://doi.org/10.1016/j.jclinepi.2011.08.002> PMID: [22075111](#)
49. Esposito D, Migliaccio-Walle K, Molsen E. Reliability and Validity of Data Sources for Outcomes Research & Disease and Health Management Programs. Lawrenceville, NJ: ISPOR, 2013–467 p. 2013.