2017

Prediction of recurrent Clostridium difficile infection using comprehensive electronic medical records in an integrated healthcare delivery system

Gabriel J. Escobar
Kaiser Permanente Division of Research

Jennifer M. Baker
Contra Costa Public Health Clinic Services

Patricia Kipnis
Kaiser Permanente Northern California

John D. Greene
Kaiser Permanente Division of Research

T. Christopher Mast
Merck Research Laboratories

See next page for additional authors

Authors

Gabriel J. Escobar, Jennifer M. Baker, Patricia Kipnis, John D. Greene, T. Christopher Mast, Swati B. Gupta, Nicole Cossrow, Vinay Mehta, Vincent Liu, and Erik R. Dubberke

ORIGINAL ARTICLE

Prediction of Recurrent *Clostridium Difficile* Infection Using Comprehensive Electronic Medical Records in an Integrated Healthcare Delivery System

Gabriel J. Escobar, MD;[1] Jennifer M. Baker, MPH, CHES;[2] Patricia Kipnis, PhD;[1,3] John D. Greene, MA;[1] T. Christopher Mast, PhD, MSc;[4] Swati B. Gupta, DrPH, MPH;[5] Nicole Cossrow, MPH, PhD;[4] Vinay Mehta, PhD;[4] Vincent Liu, MD, MS;[1,6] Erik R. Dubberke, MD[7]

BACKGROUND. Predicting recurrent *Clostridium difficile* infection (rCDI) remains difficult. METHODS. We employed a retrospective cohort design. Granular electronic medical record (EMR) data had been collected from patients hospitalized at 21 Kaiser Permanente Northern California hospitals. The derivation dataset (2007–2013) included data from 9,386 patients who experienced incident CDI (iCDI) and 1,311 who experienced their first CDI recurrences (rCDI). The validation dataset (2014) included data from 1,865 patients who experienced incident CDI and 144 who experienced rCDI. Using multiple techniques, including machine learning, we evaluated more than 150 potential predictors. Our final analyses evaluated 3 models with varying degrees of complexity and 1 previously published model.

RESULTS. Despite having a large multicenter cohort and access to granular EMR data (eg, vital signs, and laboratory test results), none of the models discriminated well (c statistics, 0.591–0.605), had good calibration, or had good explanatory power.

CONCLUSIONS. Our ability to predict rCDI remains limited. Given currently available EMR technology, improvements in prediction will require incorporating new variables because currently available data elements lack adequate explanatory power.

Infect Control Hosp Epidemiol 2017;38:1196–1203

Clostridium difficile infection (CDI) is a serious illness whose presentation can range from loose stools to profuse watery diarrhea, leading to dehydration, life-threatening complications, and sometimes death. This illness is associated with substantial morbidity, mortality, excess health services utilization, and increased cost.[1–3] The Centers for Disease Control and prevention estimated that there were 453,000 cases of incident CDI (iCDI) in 2011, with 29,000 associated deaths and 83,000 first recurrences (rCDI).[1] Recurrences are common due to persistent or newly acquired bacterial spores.[4] After initial treatment and resolution of diarrhea, up to 35% of CDI patients experience rCDI.[1,5,6] Of those with a primary recurrence, 40% will have another CDI episode, and after 2 recurrences, the likelihood of an additional episode increases to as high as 65%.[7] However, due to recent advances, this estimate may be overstated.[8,9]

Prevention of rCDI remains a critical unmet medical need, and it is desirable to predict which patients are at highest risk of recurrence. A number of research teams have developed predictive models for rCDI.[10–13] These models have had limited sample size, have been restricted to data from a single center, have employed imprecise proxies for measures of disease severity, and have made limited use of electronic medical record (EMR) data.

A need exists for risk prediction models to address these gaps. As more healthcare systems in the United States transition to fully automated EMRs, it is important to take advantage of the increased granular clinical data that are becoming available. Although health systems are beginning to experiment with predictive models embedded in EMRs,[14–16] access to such capability remains limited. The overall incidence of CDI is affected by local factors such as antimicrobial stewardship efforts, patient case mix, varying antibiotic utilization patterns, *C. difficile* strain epidemiology, and prevention. Thus, models may not be completely generalizable and may need periodic updating. Although considerable interest in predicting rCDI exists, descriptions of the performance characteristics of existing models have been limited, and few have been

Affiliations: 1. Kaiser Permanente Division of Research, Oakland, California; 2. Contra Costa Public Health Clinic Services, Martinez, California; 3. Kaiser Permanente Northern California, Oakland, California; 4. Merck Research Laboratories, North Wales, Pennsylvania; 5. Merck Vaccines, West Point, Pennsylvania; 6. Santa Clara Medical Center and Medical Offices, Kaiser Permanente Northern California, Santa Clara, California; 7. Washington University School of Medicine, St Louis, Missouri.

Received February 23, 2017; accepted July 14, 2017; electronically published August 24, 2017

 0899-823X/2017/3810-0009. DOI: 10.1017/ice.2017.176

sufficiently validated outside the populations in which they were developed. Now that treatments are available to prevent the recurrence of CDI (eg, fidaxomicin,[17,18] bezlotoxumab[19]), it is advantageous to patients and healthcare providers to identify those at greatest risk for recurrence who may benefit from the most appropriate treatments.

To address these gaps, we developed and validated rCDI predictive models in a large and representative sample of adults. For our defined population, cared for by a single medical group within an integrated delivery system, Kaiser Permanente Northern California (KPNC), comprehensive EMR data were available. Our modeling process included comparing different models and externally validating a previously published model.

MATERIALS AND METHODS

This project was approved by the KPNC Institutional Review Board for the Protection of Human Subjects, which has jurisdiction over all the hospitals and clinics described in this report.

Our setting consisted of 21 KPNC hospitals described previously.[20–22] Under a mutual exclusivity arrangement, salaried physicians of The Permanente Medical Group care for 4.2 million Kaiser Foundation Health Plan members at facilities owned by Kaiser Foundation Hospitals. All KPNC facilities (21 hospitals and an additional 60 clinics) employ the same information systems with a common medical record number.[23] Comprehensive KPNC information systems permit tracking of patient information across the continuum of care, including some aspects of care outside KPNC.[22,23] Deployment of the Epic EMR system (www.epicsystems.com), known internally as KP HealthConnect (KPHC), began in 2006 and was completed in 2010.

The eligible population (denominator) included adults ≥18 years of age with at least 1 positive test (the index test) for *C. difficile* toxins or DNA associated with a hospitalization between 2007 and 2014. The date-time stamp of the physician order for the index test was time zero (T_0) for all study measurements. Details on KPNC assays and testing procedures are provided in the Appendix.

Measures

Primary study outcome. The dependent variable was rCDI, which could occur either in the inpatient or outpatient setting. To ensure that we distinguished between incident and recurrent episodes, T_0 had to be preceded by an 84-day period with no evidence of CDI (Figure 1). A patient's treatment period extended from the first known instance of antibiotic treatment to 48 hours after conclusion of such treatment. A positive test defining a patient as having rCDI had to occur within 84 days after the end of the treatment period. Tests that occurred within the treatment period were not included. Figure 1 also shows that predictors were included if available up to 4 days after T_0, a clinically reasonable period for acquisition of information following a CDI testing order.

Mortality. We ascertained mortality using KPNC patient demographic databases and publicly available files of deceased patients provided by the Social Security Administration, as described previously.[22]

Model development. We assessed more than 150 potential predictors, including age, sex, and different configurations of historical variables (eg, antibiotic exposure, recent hospitalizations, and surgery). The final set of 23 predictors incorporated in the 3 models was based on clinical grounds, statistical performance, data abstraction burden in settings without EMRs, and (for the fully automated models) current KPNC data availability.[15,16]

Predictors fell into the following categories: demographic (age, sex), location of iCDI onset (either the inpatient setting or a skilled nursing facility), medication exposure (antibiotics, proton pump inhibitors), comorbidities (both as individual predictors as well as composite indices such as the Charlson comorbidity index[24] and the 12-month longitudinal COmorbidity Point Score, version 2, or COPS2[22]), medical history (eg, recent surgery involving the gastrointestinal tract), and physiologic markers (ie, laboratory tests, vital signs, and a severity of illness score, the Laboratory-based Acute Physiology Score, version 2 (LAPS2).[22] The LAPS2 employs 16 laboratory tests, vital signs, pulse oximetry, and neurological status checks. We categorized 24 antibiotics as

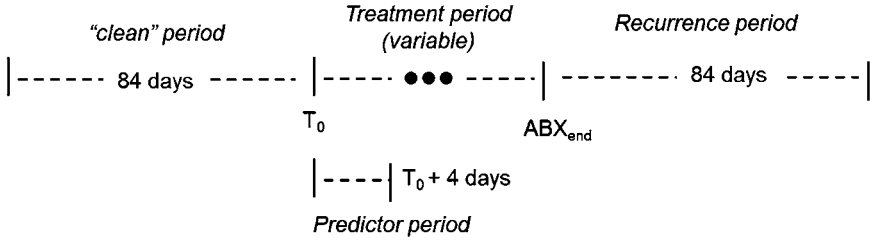

FIGURE 1. Time periods employed to define patient inclusion in cohort and patient data in predictive models. The T_0 is defined by the date/time stamp of the physician order for the index test. In order for the patient to be included in the cohort, the T_0 had to be preceded by 84 days with no positive test for *Clostridium difficile ("clean" period).* To be considered an outcome, an infection had to occur during the *Recurrence period.* This meant that a positive test result occurred within 84 days following the end of a variable treatment period (time between the T_0 and completion of antibiotic treatment, ABX_{END}). Patient data included in the predictive models had to be available within 4 days from the T_0 *(Predictor period).* See text for additional details.

high risk (eg, ciprofloxacin, clindamycin, and amoxicillin).[25–27] A full list of the predictors examined is provided in the Appendix.

Based on statistical performance, the 3 best-performing models are described here: basic, enhanced, and automated. The basic model is a parsimonious model with components that could be easily populated in most medical settings. The enhanced model is a variant of the basic model to which a limited set of variables, which could be extracted from an EMR, were added. These variables, which are part of the LAPS2 severity of illness score,[22] were based on their statistical contribution using methods described below. Finally, the automated model is based on variables that could be generated in real time given existing systems in place in KPNC.[16]

We elected to compare these final 3 models against the Zilberberg model, a previously published model by Zilberberg et al,[25] because it was based on a large cohort and the authors provided substantive detail on its statistical performance. For the Zilberberg model, we structured predictors to match the specifications of Zilberberg et al exactly. However, we did not employ their original coefficients, instead allowing these to emerge given

TABLE 1. Predictors Used Within Each Model

	Model[a]			
Predictor	Basic	Zilberberg[b]	Enhanced	Automated
Age (continuous)	x		x	
Age (splines)				x
Gastrointestinal surgery within 30 d prior to T_0	x		x	
Immunosuppression status[c]	x		x	
Locus of iCDI onset[d]	x	x	x	
Admitted from a skilled nursing facility	x		x	
≥ 2 hospitalizations within 60 d prior to T_0		x		
New gastric acid suppression (PPI) at the onset of iCDI		x		
High-risk antibiotic at the onset of iCDI[e]		x		
Fluoroquinolone at the onset of iCDI		x		
Patient in the ICU at the onset of iCDI		x		
Blood urea nitrogen			x	
Creatinine			x	
Blood urea nitrogen: creatinine			x	
Total bilirubin			x	
Arterial pH			x	
Lactate			x	
Total white blood cell count			x	
Lowest temperature within $T_0 + 4$ d			x	x
Highest temperature within $T_0 + 4$ d			x	x
LAPS2 + splines[f]				x
COPS2 + splines[g]				x
Elapsed hospital length of stay at T_0				x

NOTE. LAPS2, laboratory-based acute physiology score, version 2; COPS, comorbidity point score, version 2; T_0, Time zero (T_0) is the date-time stamp of the physician order for the index *Clostridium difficile* infection test; iCDI, incident *Clostridium difficile* infection (see text for how iCDI is defined); ICU, intensive care unit.

[a]See text for more detail on model selection.

[b]We replicated the model developed by Zilberberg et al.[25]

[c]A patient's immunosuppression status was defined using algorithmic rules using *International Classification of Disease, Ninth Revision* (ICD-9) diagnosis codes and immunocompromising medications and treatments used in the 6 mo prior to iCDI.

[d]Locus of iCDI onset is categorized as (1) community-onset, healthcare-facility–associated (iCDI diagnosed by a positive toxin test within 72 h of admission *or* iCDI diagnosed in any outpatient setting *and* a hospitalization in the prior 90 d); (2) community-onset, community-associated (reference group in model: iCDI diagnosed by a positive toxin test within 72 h of admission *or* in any outpatient setting *and* no hospitalization in the previous 90 d); or (3) hospital-onset, healthcare-facility–associated (CDI diagnosed > 72 h after hospital admission). These definitions were also used by Zilberberg et al.[25]

[e]We employed the same definitions as Zilberberg et al.[25]

[f]LAPS2 is a composite severity of illness score and employs 16 laboratory tests, vital signs, pulse oximetry, and neurological status checks.[22]

[g]COPS2 is a 12-month longitudinal comorbidity burden score that includes history elements (eg, recent surgery involving the gastrointestinal track).[22]

our population. The 4 models, arranged according to increasing complexity, are summarized in Table 1.

Statistical Methods

We divided cohort data into derivation (patients with iCDI between 2007 and 2013) and validation (iCDI in 2014) datasets. All analyses during model development were performed using the derivation dataset, with final coefficients applied once to the validation dataset. As a further precaution against overfitting, we divided derivation data into Derivation 1 (iCDI dates 2007–2012) and Derivation 2 (2013) datasets.[28] Within the Derivation 1 dataset, we identified a set of candidate predictors by first performing univariate and bivariate analyses and then applying a random forest algorithm.[28,29] We evaluated the performance and robustness of all models on the Derivation 2 data set using 5-fold cross-validation.[30] We excluded multiple models because, although they performed well in the derivation dataset, performance deteriorated dramatically following cross-validation. This was particularly true with respect to models that incorporated multiple interaction terms.

We fit a simple logistic regression, excluding deaths prior to rCDI for the basic, the enhanced, and the automated models. However, because patients with CDI have a substantial mortality risk and might die prior to developing rCDI, we evaluated several models (based on the enhanced model predictors) to address the possible impact of mortality on rCDI prediction. These included competing risk discrete survival models[29] and Cox competing risk survival regression.[31] We conducted sensitivity analyses in which we first assigned a probability of rCDI to all patients in a randomly selected portion of the derivation dataset. We then tested various models using the remaining records in which the dependent variable was not dichotomous but continuous (ie, patients who died were assigned a probability of rCDI, and then we modeled for rCDI as a continuous outcome), and we incorporated the conditional probability of mortality into the analyses. Additional details are provided in the Appendix.

We compared the discrimination of each model using the c statistic (area under the receiver operator characteristic curve),[32] calibration through calibration plots,[33] the incremental contribution of additional predictors using integrated discrimination improvement (IDI), and net reclassification improvement as recommended by Cook[34] and Pencina et al.[35] As recommended by Cook,[34] we also included the Nagelkerke pseudo-R^2 in our assessments of model performance. In standard linear regression models, the ratio of the mean-squared error to the variance of the dependent variable can be subtracted from 1 to define an R^2 that is always between 0 and 1. In a validation sample, however, the mean-squared error may exceed the variance of the dependent variable, and the resulting R^2 may be negative. A negative R^2 indicates a very poor fit with the validation sample.[36]

We also conducted sensitivity analyses in which we employed a 30-day (as opposed to an 84-day) period for outcome ascertainment.

RESULTS

We scanned KPNC databases from 2007 to 2014 and identified a total of 41,499 positive tests for *Clostridium difficile*. A total of 11,251 patients who experienced iCDI. In the derivation dataset, a total of 9,386 patients with iCDI experienced 1,311 first recurrences (14.0%); 2,197 (23.4%) patients died prior to the end of the follow-up period; and 260 (2.8%) died following a recurrence. The corresponding numbers in the validation dataset were 1,865 iCDIs, 144 (7.7%) rCDIs, 376 (20.2%) deaths prior to the end of the follow-up period, and 27 (1.4%) deaths following rCDI. The Appendix provides a flow chart describing the cohort assembly. Excluding patients who died prior to the end of the follow-up period, Table 2 summarizes our cohort characteristics, which are fairly similar to the cohort described by Zilberberg et al.[25] However, in general, the KPNC cohort was older but healthier (eg, the proportion with Charlson scores <3 was 80%, while that in the Zilberberg et al cohort was ~55%). Furthermore, the KPNC cohort generally had lower risk (eg, only 24% were receiving high-risk antibiotics, compared to 40% in the Zilberberg cohort). Expanded versions of this table are provided in the Appendix.

We compared performance of the discrete time survival and competing risk Cox regression models against the simple logistic regression algorithm where we excluded patients who died prior to an rCDI. The simple logistic regression basic, enhanced, and automated models showed performance comparable to that of the competing risk survival models.

Table 3 summarizes performance characteristics of our models in the validation dataset. All models demonstrated modest discrimination, as shown by their areas under the receiver-operator characteristic curve, or c statistics (range, 0.591–0.605) and poor explanatory power, with negative Nagelkerke pseudo-R^2s (−0.1033 to −0.0875). At a predicted risk of ≥15% the positive predictive value ranged from 11.0% to 12.1%; sensitivity ranged from 69.4% to 79.2%; and specificity ranged from 32.0% to 43.6% across the models. With this threshold, the number of patients needed to evaluate (NNE) to detect 1 case of rCDI ranged from 8.3 to 9.0 across the models. Figure 2 shows calibration of the Zilberberg model and the enhanced model; neither model was well calibrated.

Sensitivity analyses of the possible impact of mortality indicate that consideration of this issue (eg, by assigning a weighted probability of rCDI to patients who died and then modeling for rCDI as a continuous outcome) did not improve prediction. Sensitivity analyses using a 30-day (instead of 84-day) follow-up period resulted in worse model performance. Additional results are provided in the Appendix.

DISCUSSION

Using a large recent cohort, we developed and validated 3 rCDI predictive models using contemporary modeling techniques and EMR data. We also validated a previously published model[25] in a different population. However, despite including

TABLE 2. Incident *Clostridium difficile* (iCDI) Cohort Description

	Recurrence[a]	No Recurrence[a]	Total	
No.	1,455	7,223	8,678	*P* Value
Age, median y (mean ± standard deviation)	74.0 (71.3 ± 15.4)	69.0 (66.8 ± 17.2)	70.0 (67.5 ± 17.0)	<.0001
Female, No. (%)	831 (57.1)	4,131 (57.2)	4,962 (57.2)	.9557
Non-white race, No. (%)	351 (24.1)	2,068 (28.6)	2,419 (27.9)	.0005
Charlson score[b]				
0–2, No. (%)	1,141 (78.4)	5,817 (80.5)	6,958 (80.2)	.0413
3–5, No. (%)	308 (21.2)	1,394 (19.3)	1,702 (19.6)	
6 + , No. (%)	6 (0.4)	12 (0.2)	18 (0.2)	
Community onset, community associated, No. (%)[c]	315 (21.6)	2,244 (31.1)	2,559 (29.5)	<.0001
Community onset, healthcare-facility associated, No. (%)[c]	766 (52.6)	2,804 (38.8)	3,570 (41.1)	<.0001
Hospital onset, healthcare-facility associated, No. (%)[c]	374 (25.7)	2,175 (30.1)	2,549 (29.4)	.0008
No. of inpatient stays in 60 d preceding iCDI				
0, No. (%)	581 (39.9)	3,953 (54.7)	4,534 (52.2)	<.0001
1, No. (%)	603 (41.4)	2,252 (31.2)	2,855 (32.9)	
2 + , No. (%)	271 (18.6)	1,018 (14.1)	1,289 (14.9)	
Patient in intensive care at iCDI onset, No. (%)	121 (8.3)	853 (11.8)	974 (11.2)	.0005
Any antibiotics[e] at iCDI onset, No. (%)	610 (41.9)	3,170 (43.9)	3,780 (43.6)	.1682
High-risk antibiotics[e] at iCDI onset, No. (%)	351 (24.1)	1,764 (24.4)	2,115 (24.4)	.8090
Fluoroquinolone at iCDI onset, No. (%)	168 (11.5)	687 (9.5)	855 (9.9)	.0175
Low-risk antibiotics[d] at iCDI onset, No. (%)	127 (8.7)	835 (11.6)	962 (11.1)	.0017
Intravenous vancomycin at iCDI onset, No. (%)	47 (3.2)	166 (2.3)	213 (2.5)	.0361
LAPS2[e] at iCDI onset (mean ± standard deviation)	75.0 (80.8 ± 43.4)	76.0 (81.7 ± 45.3)	76.0 (81.6 ± 45.0)	.4591
COPS2[e] at iCDI onset (mean ± standard deviation)	58.0 (69.0 ± 54.0)	45.0 (60.0 ± 52.6)	48.0 (61.5 ± 52.9)	<.0001
Admitted from skilled nursing facility, No. (%)	260 (17.9)	843 (11.7)	1,103 (12.7)	<.0001

NOTE. iCDI, incident *Clostridium difficile* infection; LAPS2, laboratory-based acute physiology score, version 2; COPS2, comorbidity point score, version 2.

[a]Cohort consists of patients with iCDI. Patients who died during the follow-up period were removed from analysis.

[b]See Deyo et al[24] for details on how this score was assigned.

[c]Locus of iCDI onset is categorized as (1) community onset, healthcare-facility associated (iCDI diagnosed by a positive toxin test within 72 h of admission or iCDI diagnosed in any outpatient setting *and* a hospitalization in the prior 90 d); (2) community onset, community associated (reference group in model: iCDI diagnosed by a positive toxin test within 72 h of admission or in any outpatient setting *and* no hospitalization in the previous 90 d); or (3) hospital onset, healthcare-facility associated (CDI diagnosed >72 h after hospital admission). These definitions were also used by Zilberberg et al.[25]

[d]We employed the same antibiotic classifications as Zilberberg et al.[25]

[e]For an extended definition of LAPS2 and (COPS2), refer to the text and Escobar et al.[22] For both of these scores, increasing values are associated with increasing mortality risk. The univariate relationship of an admission LAPS2 with 30-d mortality is as follows: 0–59, 1.0%; 60–109, 5.0%, 110 + , 13.7%; the univariate relationship of COPS2 with 30-d mortality is as follows: 0–39, 1.7%; 40–64, 5.2%, 65 + , 9.0%.

TABLE 3. Model Performance in the Validation Dataset[a] at a Predicted Risk of ≥15%

Model[b,c]	c Statistic	R^2	Brier Score	Sensitivity	Specificity	PPV	NPV	NNE	NRI	IDI
Age ≥65 years	0.546	−0.1131	0.0944	67.36	41.86	11.04	92.30	9.06	…	…
Basic model	0.591	−0.0910	0.0937	75.69	41.19	12.11	94.06	8.26	0.0766	0.011
Zilberberg model	0.591	−0.0875	0.0933	74.31	39.03	11.54	93.42	8.66	0.0412	0.011
Enhanced model	0.587	−0.0924	0.0936	69.44	43.64	11.66	93.03	8.58	0.0387	0.014
Automated model	0.605	−0.1033	0.0942	79.17	32.04	11.09	93.49	9.02	0.0199	0.012

NOTE. c statistic, area under the receiver operator characteristic curve; R^2, Nagelkerke's pseudo-R^2; PPV, positive predictive value; NPV, negative predictive value; NNE, number of incident cases one would need to evaluate to detect one recurrence; NRI, net reclassification improvement; IDI, integrated discrimination improvement; iCDI, incident *Clostridium difficile* infection.

[a]The validation dataset consisted of 1,865 iCDI patients, of whom 144 developed rCDI. A total of 376 iCDI patients died (and thus could not be assessed for recurrence).

[b]See text for description of the 4 models. "Age ≥65 years" refers to a simple decision rule based on age alone. Sensitivity, PPV, NPV, NNE, NRI, and IDI are based on the model giving a predicted recurrence risk of ≥15% within 84 days.

[c]We conducted sensitivity analyses using predicted risk of ≥20%, ≥25%, and ≥30%. These results are provided in the Appendix.

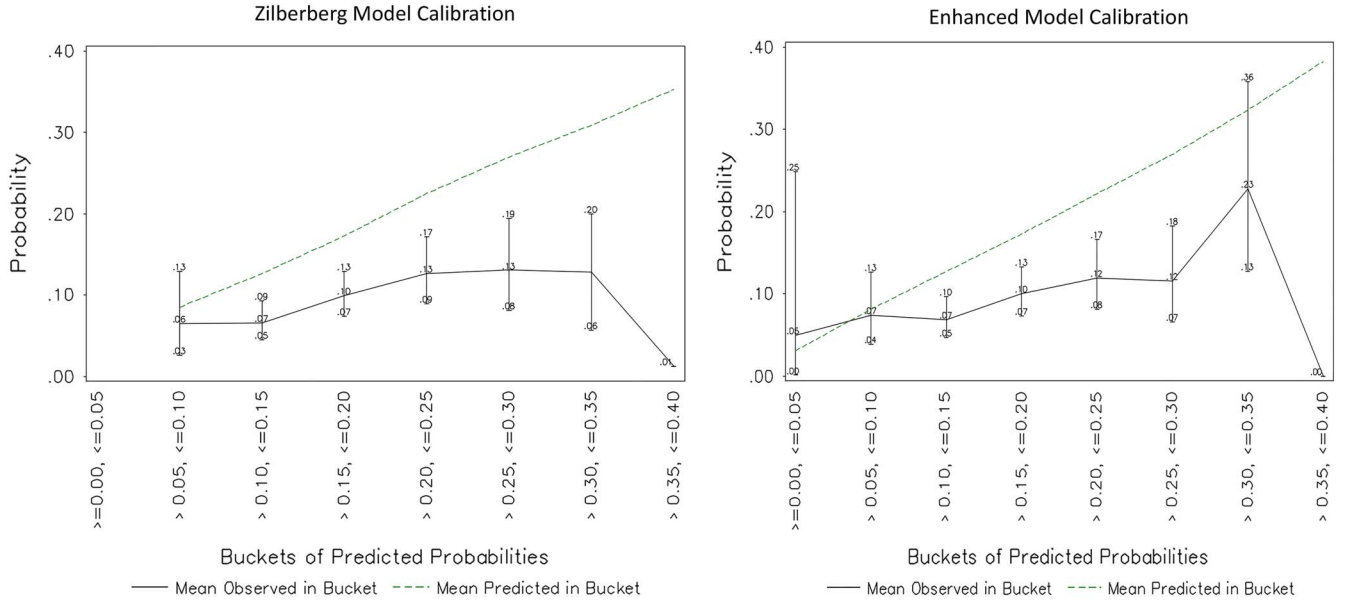

FIGURE 2. Model Calibration Using the Validation Dataset. For both plots, the X axis shows predicted rates of recurrent CDI in 5% increments, while the Y axis shows the actual observed rates (with associated 95% confidence intervals) in the validation dataset for all observations with that predicted level of risk. The dotted line shows what would be found were calibration to be perfect. For both the Zilberberg and Enhanced models, calibration is poor: calibration fails at levels above 10% predicted risk. Observed rates do not approach predicted rates, meaning that both models over-predict recurrent CDI. Additional calibration figures, including Hosmer-Lemeshow plots, are provided in the Appendix.

highly granular EMR data (eg, vital signs, laboratory tests, composite severity of illness scores, and longitudinal comorbidity), the models and underlying data had poor ability to predict rCDI. We formally tested a common assumption made by many investigators (ie, that deaths can simply be excluded from the numerator). We found that this approach is justified, and that including patients who die prior to the conclusion of the follow-up period did not improve prediction. Lastly, we found that shortening the length of follow-up to 30 days resulted in worse model performance.

Some authors have reported better model performance. Examination of these other studies paints a less optimistic picture. Hu et al[10] report the use of machine-learning approaches and a c statistic of 0.80 in their validation dataset. However, this study had a very small sample size (N = 110, with N = 64 in the validation dataset) and did not employ cross-validation (ie, no formal assessment of the possibility that model performance in a different population might be poor). We were able to achieve c statistics that were this high in our derivation dataset, but these apparently successful models demonstrated considerable instability during cross-validation. We did not pursue them further and chose more parsimonious models.

Contrary to previous literature reports, some predictors (eg, specific antibiotic exposures) were of limited value, particularly in models that included severity of illness. This probably reflects the fact that severity of illness is highly correlated (and may, in fact, be the underlying risk factor) with other predictors (eg, intensive care and antibiotics known to predispose for CDI). We deliberately focused on predicting rCDI in iCDI cases, though previous CDI is a well-known risk factor for recurrence. It is possible that, had we included prior CDI as a predictor, we might have achieved better model performance. However, models that included the COPS2 score (a longitudinal comorbidity measure that captures information from the preceding 12 months) did not perform much better.

Multiple investigators, using a variety of statistical approaches, including machine-learning methods, have been unable to produce static models with better performance using the currently available set of predictors. While it is true that many predictors reach statistical significance in bivariate analyses (particularly when the sample size is large), the clinical significance may be muted because the relative proportions of patients with and without recurrence are not that different. Further, it is clear that the risk factors (age, antibiotic exposure, severity of illness) that place an individual at risk for iCDI are also risk factors for rCDI. Thus, future efforts ought to be placed on identifying better predictors rather than on using different statistical approaches with the currently available predictors. New predictors may include newer biomarkers (eg, indicators of underlying predisposition to recurrence), environmental factors (eg, proximity to other CDI patients, presence of *C. difficile* spores[4]), behavioral aspects (eg, handwashing), and/or molecular markers (eg, information on specific *C. difficile* strains). It is also important to consider rCDI in an ecological context, and future predictive models may need to be explicit about including environmental and ecological predictors

(eg, isolation rooms, who is roomed where, other family members exposure), if such data become available.

One alternative that we did not explore because it is currently not feasible with existing EMRs, was to develop dynamic models. In contrast to the static approach we and others have employed (ie, providing a single probability estimate based on a discrete set of predictors available at some T_0), such models adjust posterior probabilities based on new information. In the case of rCDI, having additional information on both antibiotic treatment as well as other exposures (eg, proton pump inhibitors) could have dramatic effects on our ability to predict recurrence.[37,38] The development of such models would require EMRs with greater capabilities than those currently available.

Our study had several additional limitations. Due to resource limitations and sparse data, we limited our cases to inpatient iCDI. During this study, KPNC implemented aggressive efforts to reduce CDI. As a result, our data show that the incidences of iCDI and rCDI were decreasing in our study cohort. Despite these limitations, models to predict recurrence have value. They do permit identification of patient subsets with elevated or very low risk. In some scenarios, and in the context of discrete interventions, the use of these models might improve outcomes and decrease costs. In addition, existing models point to predictors that can be assessed in the future, such as the aforementioned ecological ones.

Compared to our ability to predict other outcomes (eg, death, unplanned transfer to intensive care),[16,20,22] our ability to predict rCDI is limited and contrasts with much better ability to predict iCDI.[39,40] Given the major consequences of rCDI on patient outcomes, our results support the need to expand research on the prevention and treatment of recurrence. Such research may also result in the identification of novel predictors that are currently unavailable even in the most comprehensive EMRs.

ACKNOWLEDGMENTS

This project was funded by a grant from Merck Sharp & Dohme Corporation, Whitehouse Station, New Jersey. The authors wish to thank Juan Carlos LaGuardia for help assembling the dataset, Dr Tracy Lieu for reviewing the manuscript, Vanessa Rodriguez for formatting the text for publication, Anna Cardellino for her assistance in drafting the protocol, and Mary Beth Dorr for her review and guidance in the analysis.

Financial support: Dr Vincent Liu was funded by a National Institutes of Health award (grant no. K23GM112018).

Potential conflicts of interest: The Kaiser Permanente authors Escobar, Kipnis, Liu, Greene, and Baker have no conflicts of interest to report. Dr Erik Dubberke has received grant support from Rebiotix, Merck, and Sanofi Pasteur; he also has consulting and advisory board relationships with Rebiotix, Summit, GSK, Valenva, Sanofi Pasteur. The remaining coauthors Cossrow, Gupta, Mast, and Mehta are or were employees of Merck Sharp & Dohme Corporation, a subsidiary of Merck & Co., Kenilworth, New Jersey, and potentially own stock and/or hold stock options in the company.

Address correspondence to Gabriel J. Escobar, MD, Systems Research Initiative, Kaiser Permanente Division of Research, 2000 Broadway Ave (032 R01), Oakland, CA 94612-2304 (gabriel.escobar@kp.org) or John Greene, MA, Systems Research Initiative, Kaiser Permanente Northern California Division of Research, 2000 Broadway Ave, Oakland, CA 94612 (john.d.greene@kp.org).

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/ice.2017.176

REFERENCES

1. Lessa FC, Winston LG, McDonald LC. Burden of *Clostridium difficile* infection in the United States. *N Engl J Med* 2015;372:2369–2370.
2. Kwon JH, Olsen MA, Dubberke ER. The morbidity, mortality, and costs associated with *Clostridium difficile* infection. *Infect Dis Clin North Am* 2015;29:123–134.
3. Olsen MA, Young-Xu Y, Stwalley D, et al. The burden of *Clostridium difficile* infection: estimates of the incidence of CDI from US administrative databases. *BMC Infect Dis* 2016;16:177.
4. Freedberg DE, Salmasian H, Cohen B, Abrams JA, Larson EL. Receipt of antibiotics in hospitalized patients and risk for *Clostridium difficile* infection in subsequent patients who occupy the same bed. *JAMA Intern Med* 2016;176:1801–1808.
5. McFarland LV. Renewed interest in a difficult disease: *Clostridium difficile* infections—epidemiology and current treatment strategies. *Curr Opin Gastroenterol* 2009;25:24–35.
6. Bouza E. Consequences of *Clostridium difficile* infection: understanding the healthcare burden. *Clin Microbiol Infect* 2012;18(Suppl 6):5–12.
7. McFarland LV, Elmer GW, Surawicz CM. Breaking the cycle: treatment strategies for 163 cases of recurrent *Clostridium difficile* disease. *Am J Gastroenterol* 2002;97:1769–1775.
8. Zilberberg MD, Reske K, Olsen M, Yan Y, Dubberke ER. Risk factors for recurrent *Clostridium difficile* infection (CDI) hospitalization among hospitalized patients with an initial CDI episode: a retrospective cohort study. *BMC Infect Dis* 2014;14:306.
9. Sheitoyan-Pesant C, Abou Chakra CN, Pepin J, Marcil-Heguy A, Nault V, Valiquette L. Clinical and healthcare burden of multiple recurrences of *Clostridium difficile* infection. *Clin Infect Dis* 2016;62:574–580.
10. Hu MY, Katchar K, Kyne L, et al. Prospective derivation and validation of a clinical prediction rule for recurrent *Clostridium difficile* infection. *Gastroenterology* 2009;136:1206–1214.
11. Eyre DW, Walker AS, Wyllie D, et al. Predictors of first recurrence of *Clostridium difficile* infection: implications for initial management. *Clin Infect Dis* 2012;55:S77–S87.
12. Hebert C, Du H, Peterson LR, Robicsek A. Electronic health record–based detection of risk factors for *Clostridium difficile* infection relapse. *Infect Control Hosp Epidemiol* 2013;34:407–414.
13. D'Agostino RB Sr., Collins SH, Pencina KM, Kean Y, Gorbach S. Risk estimation for recurrent *Clostridium difficile* infection based on clinical factors. *Clin Infect Dis* 2014;58:1386–1393.

14. Kollef MH, Chen Y, Heard K, et al. A randomized trial of real-time automated clinical deterioration alerts sent to a rapid response team. *J Hosp Med* 2014;9:424–429.
15. Escobar GJ, Dellinger RP. Early detection, prevention, and mitigation of critical illness outside intensive care settings. *J Hosp Med* 2016;11:S5–S10.
16. Escobar GJ, Turk BJ, Ragins A, et al. Piloting electronic medical record-based early detection of inpatient deterioration in community hospitals. *J Hosp Med* 2016;11:S18–S24.
17. Watt M, Dinh A, Le Monnier A, Tilleul P. Cost-effectiveness analysis on the use of fidaxomicin and vancomycin to treat *Clostridium difficile* infection in France. *J Med Econ* 2017;20:678–686.
18. Nelson RL, Suda KJ, Evans CT. Antibiotic treatment for *Clostridium difficile*–associated diarrhoea in adults. *Cochrane Database Syst Rev* 2017;3:CD004610.
19. Wilcox MH, Gerding DN, Poxton IR, et al. Bezlotoxumab for prevention of recurrent *Clostridium difficile* infection. *N Eng. J Med* 2017;376:305–317.
20. Escobar GJ, Greene JD, Gardner MN, Marelich GP, Quick B, Kipnis P. Intra-hospital transfers to a higher level of care: contribution to total hospital and intensive care unit (ICU) mortality and length of stay (LOS). *J Hosp Med* 2011;6:74–80.
21. Liu V, Kipnis P, Rizk NW, Escobar GJ. Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system. *J Hosp Med* 2012;7:224–230.
22. Escobar GJ, Gardner MN, Greene JD, Draper D, Kipnis P. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Med Care* 2013;51:446–453.
23. Selby JV. Linking automated databases for research in managed care settings. *Ann Intern Med* 1997;127:719–724.
24. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 1992;45:613–619.
25. Zilberberg MD, Reske K, Olsen M, Yan Y, Dubberke ER. Development and validation of a recurrent *Clostridium difficile* risk-prediction model. *J Hosp Med* 2014;9:418–423.
26. Dubberke ER, Reske KA, Yan Y, Olsen MA, McDonald LC, Fraser VJ. *Clostridium difficile*–associated disease in a setting of endemicity: identification of novel risk factors. *Clin Infect Dis* 2007;45:1543–1549.
27. Dubberke ER, Yan Y, Reske KA, et al. Development and validation of a *Clostridium difficile* infection risk prediction model. *Infect Control Hosp Epidemiol* 2011;32:360–366.
28. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer Verlag; 2009.
29. Allison PD. *Logistic Regression Using SAS: Theory and Application*. 2nd ed. Cary, NC: SAS Institute; 2012.
30. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. *Mathemat Intelligenc* 2005;27:83–85.
31. Hosmer DW, Lemeshow S. *Applied Survival Analysis: Regression Modelling of Time to Event Data*. Hoboken, NJ: Wiley; 2008.
32. Cook DA, Duke G, Hart GK, Pilcher D, Mullany D. Review of the application of risk-adjusted charts to analyse mortality outcomes in critical care. *Crit Care Resusc* 2008;10:239–251.
33. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. *Stat Method Med Res* 2014;25:1692–1706.
34. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115:928–935.
35. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–172; discussion 207–212.
36. Estrella A, Mishkin FS. Predicting US recessions: financial variables as leading indicators. *Rev Econ Statist* 1998;80:45–61.
37. McDonald EG, Milligan J, Frenette C, Lee TC. Continuous proton pump inhibitor therapy and the associated risk of recurrent *Clostridium difficile* Infection. *JAMA Intern Med* 2015;175:784–791.
38. Deshpande A, Pasupuleti V, Thota P, et al. Risk factors for recurrent *Clostridium difficile* infection: a systematic review and meta-analysis. *Infect Control Hosp Epidemiol* 2015;36:452–460.
39. Kuntz JL, Johnson ES, Raebel MA, et al. Predicting the risk of *Clostridium difficile* infection following an outpatient visit: development and external validation of a pragmatic, prognostic risk score. *Clin Microbiol Infect* 2015;21:256–262.
40. Kuntz JL, Smith DH, Petrik AF, et al. Predicting the risk of *Clostridium difficile* infection upon admission: a score to identify patients for antimicrobial stewardship efforts. *Perm J* 2016;20:20–25.