Open Forum Infectious Diseases

MAJOR ARTICLE

A Risk Score to Predict *Clostridioides difficile* Infection

Laurie Aukes,[1] Bruce Fireman,[1] Edwin Lewis,[1] Julius Timbol,[1] John Hansen,[1] Holly Yu,[2] Bing Cai,[2] Elisa Gonzalez,[2] Jody Lawrence,[2] and Nicola P. Klein[1]

[1]Kaiser Permanente Vaccine Study Center, Oakland, California, USA, [2]Pfizer, Inc., Collegeville, Pennsylvania, USA

Background. *Clostridioides difficile* infection (CDI) is a major cause of severe diarrhea. In this retrospective study, we identified CDI risk factors by comparing demographic and clinical characteristics for Kaiser Permanente Northern California members ≥18 years old with and without laboratory-confirmed incident CDI.

Methods. We included these risk factors in logistic regression models to develop 2 risk scores that predict future CDI after an Index Date for Risk Score Assessment (IDRSA), marking the beginning of a period for which we estimated CDI risk.

Results. During May 2011 to July 2014, we included 9986 CDI cases and 2 230 354 members without CDI. The CDI cases tended to be older, female, white race, and have more hospitalizations, emergency department and office visits, skilled nursing facility stays, antibiotic and proton pump inhibitor use, and specific comorbidities. Using hospital discharge as the IDRSA, our risk score model yielded excellent performance in predicting the likelihood of developing CDI in the subsequent 31–365 days (C-statistic of 0.848). Using a random date as the IDRSA, our model also predicted CDI risk in the subsequent 31–365 days reasonably well (C–statistic 0.722).

Conclusions. These results can be used to identify high-risk populations for enrollment in *C difficile* vaccine trials and facilitate study feasibility regarding sample size and time to completion.

Keywords. *Clostridioides difficile* risk score model.

Clostridioides difficile infection (CDI) accounts for a large portion of nosocomial morbidity and mortality. In 2011, CDI caused an estimated half a million infections in the United States with approximately 29 000 associated deaths [1]. In general, the incidence and virulence of CDI have been increasing. Hospitalizations due to CDI increased from 8.8 per 1000 nonpregnant adults in 2004 to 13.7 in 2013 [2]. Antibiotics, hospitalization, and older age are important risk factors [3, 4] which have been associated with this increase; however, CDI also occurs without these risk factors (eg, pregnant women, emergency department [ED], outpatient, and nonhealthcare settings) [4]. In contrast, limited information exists regarding the burden of CDI in specific populations such as individuals with underlying medical conditions and those receiving care in the ED and outpatient settings.

The clinical importance of *C difficile* has resulted in it being targeted for vaccine development [5–8]. However, predicting who is most likely to develop CDI within the next year and be eligible for *C difficile* vaccine clinical studies remains a challenge. A risk score model that incorporates important risk factors from inpatient and outpatient settings to predict incident CDI could potentially identify such individuals. Although risk models have been previously proposed, they have had limited generalizability due to small samples, inclusion of data from only 1 healthcare setting (eg, inpatient) or age group (eg, ≥65 years), and/or were tailored to specific medical facilities [9–17]. To date, no risk score has incorporated into 1 model the many potential CDI risk factors often available in an electronic medical record (EMR).

We aimed to build CDI risk score models that may be more generalizable to other populations and healthcare settings that could be used to target high-risk individuals for studying preventive measures including *C difficile* vaccination. To build this risk score, we identified risk factors associated with CDI at Kaiser Permanente Northern California (KPNC) and predicted the risk of CDI.

METHODS

Setting

Kaiser Permanente Northern California delivers integrated healthcare services to members who receive almost all of their care at KPNC facilities, which includes 65 medical clinics and 27 hospitals. At the time of the study, this included 3.2 million members, including 2.45 million adults aged ≥18 years. Kaiser Permanente Northern California's EMR captures all healthcare encounters including diagnoses, medications, laboratory tests, and any CDI event. Kaiser Permanente Northern California's single centralized laboratory conducted *C difficile* stool testing by 2-stage procedure: (1) enzyme immunoassay for

Received 20 October 2020; editorial decision 26 January 2021; accepted 31 January 2021.

Correspondence: L. Aukes, RN, CCRA, Kaiser Permanente Vaccine Study Center, 1 Kaiser Plaza, 16th Floor, Oakland, CA 94612 (laurie.a.aukes@kp.org).

Open Forum Infectious Diseases®2021

DOI: 10.1093/ofid/ofab052

C difficile antigen, glutamate dehydrogenase (Alere, Waltham, Massachusetts); and (2) positive and equivocal enzyme immunoassay tests confirmed using polymerase chain reaction (PCR) to detect *C difficile* toxin B gene sequences (Cepheid, Sunnyvale, California). Only freshly passed stool specimens (no cathartic or enema) with liquid or loose stool is acceptable; formed stool samples are rejected. Kaiser Permanente Northern California has region-wide standardized infection control policies; however, our approximately 100 hospitals and clinics may each have facility-specific CDI risk and/or protective factors (eg, outbreaks, member social/economic make-up, facility size/layout, and staff practices) potentially mirroring variations in other healthcare settings and adding to the generalizability of our risk scores.

Study Population and Design

This retrospective study was conducted in 2 phases using KPNC's EMR from May 2010 to July 2014. In Phase 1, we ascertained potential risk factors for CDI, and, in Phase 2, we developed CDI risk scores. Using the following overall approach, the unit of study for both phases was an index date (Figure 1). In Phase 1, among all KPNC members aged ≥18 years between May 2011 and July 2014, we identified all incident CDI cases. We used the remaining KPNC population for comparison. We profiled for potential CDI risk factors by comparing demographic and clinical characteristics between those with and without CDI in the 1 year before an index date. The index date for this comparison was either the CDI diagnosis date or a randomly chosen date for the remaining population without CDI. We required 1 year of membership before the CDI diagnosis date or random date to collect characteristics.

Individuals were categorized as having incident CDI (CDI cases) or having no record of CDI (comparison population) as follows. (1) For CDI cases, we defined an incident CDI case as having a *C difficile* PCR-positive test during the study period without history of laboratory-confirmed *C difficile* during the 1 year before the CDI diagnosis date (the first *C difficile* PCR-positive test on study). If a subject had a distant history of prior CDI, the CDI event during the study period must have occurred at least 1 year after a prior positive PCR test to be considered incident CDI. (2) In the comparison population, we included all individuals without a positive PCR test for *C difficile* during the 1 year before a random date. We used the SAS uniform random number generator function to select a random date during the study period for each individual without CDI to act as an index date for risk factor comparison. Thus, for both CDI cases and comparison population, we ensured no CDI episodes occurred in the 1 year before the diagnosis or random date.

In Phase 2, we used the results from Phase 1 to inform the development of risk score models to predict CDI by anchoring the models on an Index Date for Risk Score Assessment (IDRSA). Using risk factors (identified in Phase 1) present for each individual in the year before the IDRSA, we created risk scores to predict CDI in the year after the IDRSA.

Utilizing a subset of the study period and population, we developed models to evaluate 2 different IDRSAs: (1) a hospital discharge IDRSA model (anchored on hospital discharge dates) included only members aged ≥18 years who had a hospitalization between May 2011 and July 2013; (2) a random IDRSA model (anchored on randomly selected dates, discussed further below) included all members ≥50 years between May 2011 and July 2013.

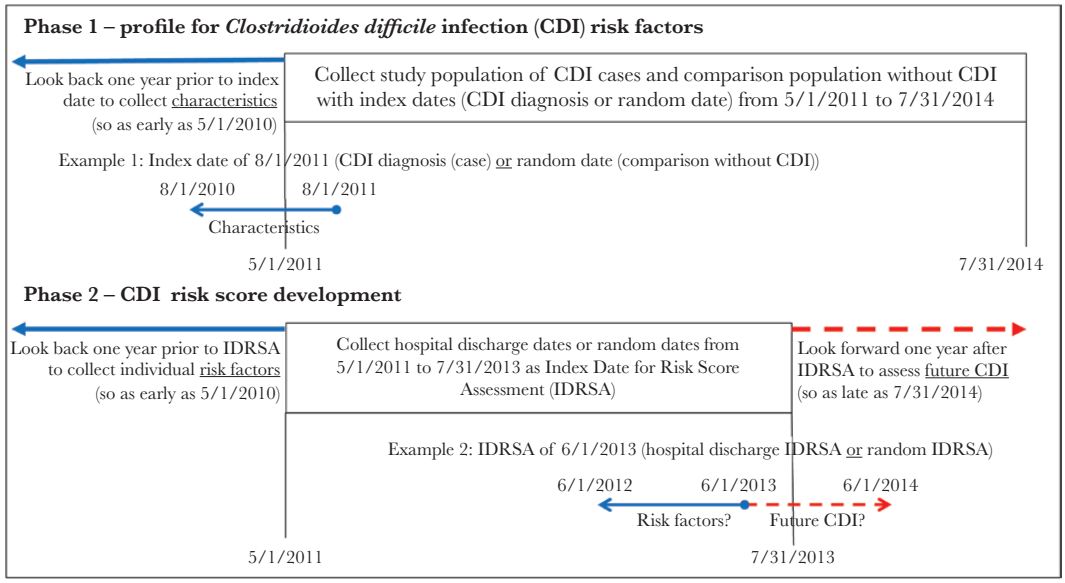

Figure 1. Study Design. Phase 1 example demonstrates that for an index date of 8/1/2011, we looked back one year to 8/1/2010 to identify and characterize potential risk factors for *Clostridioides difficile* infection (CDI) at the population level. Phase 2 example demonstrates that for an Index Date for Risk Score Assessment (IDRSA) of 6/1/2013, we looked back one year to 6/1/2012 to collect individual risk factors for CDI (as identified in Phase 1) and then looked forward one year to 6/1/2014 to predict risk for CDI during that year.

This age group was selected for the random model because those aged 50+ are more likely to develop CDI and be the target population for a *C difficile* vaccine efficacy study. For each model, we required membership at least 1 year before the IDRSA (to collect risk factors) and 1 year after the IDRSA (to look for future CDI).

Patient Consent

This study was determined exempt by the KPNC Institutional Review Board. Waivers of informed consent and Health Insurance Portability and Accountability Act (HIPAA) were granted. Kaiser Permanente Northern California's Institutional Review Board approved this study.

Statistical Analysis

In Phase 1, we compared all incident CDI cases with the comparison population. We examined covariates potentially related to CDI including demographics, comorbid medical conditions (using Healthcare Cost and Utilization Project and *International Classification of Diseases, Ninth Revision*, codes [Supplementary Table 1]), healthcare utilization (hospitalization, outpatient, ED, and long-term care), medications (antibiotics, proton-pump inhibitors, and immunocompromising medications), and Charlson comorbidity scores [18]. We assessed all risk factors for their presence during different baseline periods (eg, 12 weeks or 1 year) before the diagnosis (CDI cases) or random date (comparison population).

We calculated CDI incidence rates per 100 000 person-years and 2-sided 95% confidence intervals using exact Poisson methods. Incidence rates were calculated overall and for subgroups of interest defined by demographics and comorbidities.

In Phase 2, we used a hospital discharge date or random date as our IDRSA to mark the beginning of the period for which we estimated risk for CDI. We used logistic regression to estimate the risk of developing CDI in relation to the risk factors identified during the baseline period (eg, 12 weeks or 1 year) before the IDRSA.

We used 2 different IDRSAs to create models to predict the future risk of CDI in the 31- to 365-day period starting at each IDRSA. The primary risk score model was anchored to hospital discharge date as the IDRSA because we envisioned that hospital discharge could be a touch point to identify high-risk populations for recruitment into *C difficile* vaccine trials. Using a cross-validation strategy, we developed the hospital discharge model using a random 70% of the data and then evaluated it with the remaining 30% [19].

To achieve a broader representation of the general population and make decisions about people not in the hospital, we created a simplified secondary risk score model in which a randomly selected date (using the same SAS random number function discussed above) was used as the IDRSA. The rationale for using a random date as the IDRSA was to simulate the real-world setting whereby a nonhospitalized individual's eligibility for inclusion in *C difficile* vaccine trial could be evaluated based on the presence of risk factors.

Each model was fit to datasets created specifically for each IDRSA. (1) The hospital discharge IDRSA dataset included demographic and hospitalization variables identified as risk factors in Phase 1 (eg, age, diagnoses, medications) for all hospitalized individuals (CDI cases and comparison population) aged 18+ years; individuals not hospitalized did not contribute. To choose which risk factors to use, we started with approximately 200 possible risk factors and used a forward selection process consisting of an automated stepwise hierarchy with built-in *P* value threshold of ≤0.15 for each risk factor to be kept in the model; 100 risk factors remained that met the threshold. (2) The random IDRSA dataset included demographic and risk factor variables for all individuals (CDI cases and comparison population) aged ≥50 years. We chose approximately 20 risk factors based on known CDI risk factors (eg, age, healthcare utilization, antibiotics, medical history) that would be more generalizable and practical for use at a clinical site, including sites with less extensive EMR.

For each cohort—the cohort of hospitalized patients aged ≥18 years on their discharge date (hospital discharge IDRSA) and the cohort of members aged ≥50 years on a randomly selected date (random IDRSA)—we used logistic regression to examine the risk of CDI in relation to risk factors and demographics. For each cohort, a logistic regression model was fitted to yield a predicted probability of CDI risk for each person during each follow-up period. Each fitted model also yielded an estimate of the odds ratio (and corresponding relative risk) associated with each risk factor. We reported 95% Wald confidence intervals for the odds ratios and corresponding 2-sided *P* values.

From both risk score models, we excluded the first 30 days after IDRSA to theoretically allow for time for vaccination and development of immunity. Therefore, we conducted a supplemental analysis to create a separate random date IDRSA model with follow-up of 1–30 days after IDRSA to evaluate the number of CDI cases that might occur early and thus not be included in the models with longer follow-up. We used SAS software, version 9.2 (SAS Institute) for all analyses.

RESULTS

Phase 1 Results: Risk Factors

From May 2011 through July 2014, we identified 9986 incident CDI cases and 2 230 354 in the comparison population, for a CDI incidence rate of 141 per 100 000 person-years among members aged ≥18. Individuals with CDI were more likely to be older (aged ≥65 years, 59% vs 21%), female (61% vs 53%), and white race (70% vs 53%) than those in the comparison population (Table 1).

During the year before incident CDI, a greater percentage of people with CDI than without were hospitalized (69% vs 10%),

Table 1. Incidence of *Clostridioides difficile* Infection vs Comparison by Demographics, Individuals Aged 18+ Years, Kaiser Permanente Northern California, May 2011 to July 2014

Covariate	Number of CDI Cases	%CDI Cases	Non-CDI Population	%Non-CDI Population	Person-Years (PY) in Total Population	CDI Incidence per 100 000 PY (95% CI)
Total	9986	100.0	2 230 354	100.0	7 079 474	141 (138–144)
Age Category						
18–49 Years	1692	16.9	1 146 768	51.4%	3 629 134	47 (44–49)
50–64 Years	2359	23.6	627 212	28.1%	1 989 444	119 (114–124)
65–74 Years	2037	20.4	261 726	11.7%	833 491	244 (234–255)
75–84 Years	2300	23.0	139 632	6.3%	448 505	513 (492–534)
85+ Years	198	16.0	55 016	2.5%	178 900	893 (850–928)
All ≥50 Years	8294	83.1	1 083 586	48.6%	3 450 341	240 (235–246)
All ≥65 Years	5935	59.4	456 374	20.5%	1 460 896	406 (396–417)
Gender						
Female	6038	60.5	1 182 338	53.0%	3 755 268	161 (157–165)
Male	3948	39.5	1 048 016	47.0%	3 324 206	119 (115–123)
Race						
White	6973	69.8	1 177 719	52.8%	3 743 627	186 (182–191)
Black	917	9.2	159 662	7.2%	507 430	181 (169–193)
Asian	877	8.8	381 119	17.1%	1 207 107	73 (68–78)
Hispanic	922	9.2	267 039	12.0%	846 757	109 (102–116)

Abbreviations: CDI, *Clostridioides difficile* infection; CI, confidence interval.

had an ED visit (51% vs 14%), skilled nursing stay (25% vs 0.6%), nursing home stay (4% vs 0.1%), or at least 10 outpatient visits (53% vs 16%), used a proton pump inhibitor (36% vs 7%), had a Charlson comorbidity index score of 3 (which indicates moderate risk of death) (11% vs 2%), and had multiple medical conditions, such as pneumonia (21% vs 1%), chronic kidney disease (26% vs 4%), coronary artery disease (22% vs 3%), congestive heart failure (22% vs 2%), urinary tract infection (28% vs 3%), diabetes (23% vs 6%), and peripheral vascular disease (9% vs 1%), and were prescribed an antibiotic in the prior 12 weeks (81% vs 11%) (Table 2).

Table 2. Selected Risk Factors Within 1 Year (Except as Noted[a]) Before Diagnosis or Random Date, *Clostridioides difficile* Infection vs Comparison, Individuals Aged 18+ Years, Kaiser Permanente Northern California, May 2011 to July 2014

	CDI Cases		Non-CDI Population			
Risk Factor[a]	N = 9986	%	N = 2 230 354	%	Ratio (%CDI Cases/%Non-CDI)	CDI Incidence per 100 000 PY (95% CI)
≥1 inpatient hospitalization	6856	68.7	227 213	10.2	6.7	927 (905–949)
≥2 inpatient hospitalization	4166	41.7	62 860	2.8	14.9	1967 (1908–2028)
Emergency room visit	5119	51.3	302 069	13.5	3.8	527 (513–542)
>10 outpatient visits	5219	53.3	290 763	15.9	3.4	559 (543–573)
Skilled nursing facility stay	2529	25.3	12 373	0.6	46.1	5371 (5163–5584)
Custodial care facility stay	368	3.7	2412	0.1	33.5	4189 (3772–4640)
Systemic antibiotic use (within 12 weeks)	8039	80.5	236 476	10.6	7.6	1040 (1018–1064)
Prescription systemic proton pump inhibitor use	3541	35.5	148 700	6.7	5.3	736 (712–761)
Congestive heart failure	2197	22.0	34 842	1.6	13.8	1877 (1799–1957)
Chronic kidney disease (less severe)	2564	25.7	98 219	4.4	5.8	805 (774–837)
Severe chronic kidney disease (on dialysis)	636	6.4	5309	0.2	32.0	3386 (3128–3659)
Diabetes mellitus	2341	23.4	129 047	5.8	4.0	564 (541–587)
Chronic obstructive pulmonary disease	1750	17.5	70 504	3.2	5.5	767 (731–803)
Peripheral vascular disease	884	8.9	19 748	0.9	9.9	1356 (1268–1448)
Coronary artery disease	2231	22.3	74 364	3.3	6.8	922 (884–961)
Acute myocardial infarction	493	4.9	5882	0.3	16.3	2447 (2236–2673)
Liver disease	1447	14.5	45 955	2.1	6.9	966 (917–1017)
Pneumonia	2124	21.3	29 136	1.3	16.4	2150 (2060–2244)
Urinary tract infection	2786	27.9	57 372	2.6	10.7	1466 (1412–1521)
Charlson comorbidity score of 3	1073	10.7	48 799	2.2	4.9	681 (641–723)

Abbreviations: CDI, *Clostridioides difficile* infection; CI, confidence interval; PY, person-years.

[a]Number of patients with that risk factor (eg, comorbid condition, medication, healthcare use) within 1 year before CDI diagnosis date (first positive *C difficile* toxin text) or random date except for antibiotic use that was within 12 weeks before diagnosis or random date.

Phase 2 Results: Risk Scores

Among the 104 518 hospital discharges in the validation set (representing 30% of all hospital discharges), the hospital discharge IDRSA model yielded excellent performance in predicting the likelihood of developing CDI during the 31–365 days after hospital discharge in the 30% of data withheld for validation. A C-statistic of 0.848 was generated as the measure of fit for the model (Supplementary Figure 1).

From the model validation results based on risk factors included in the hospital discharge IDRSA model (Table 3 and Supplementary Table 2), we can see the number of hospitalized individuals aged ≥18 years expected to develop CDI by risk score threshold or range as demonstrated by the following examples. (1) Higher risk scores of ≥0.30 will target 1.1% of the hospital discharges (1146 of 104 518) and 11.1% of all posthospital CDI cases (379 of 3423) (Table 3). (2) Alternatively, lower risk scores of ≥0.05 will target 11.9% of hospital discharges (12 423 of 104 518) and 57.2% of all posthospital CDI cases (1957 of 3423) (Table 3), which would provide better representation of hospitalized adults for enrollment in large Phase 3 *C difficile* vaccine efficacy studies. (3) In the 0.30 to 0.35 risk score range, we predict 35.2% will get CDI (Supplementary Table 2).

Our random IDRSA model also predicted CDI reasonably well in the 31- to 365-day follow-up period (C–statistic 0.722) (Supplementary Figure 2). From the model validation results based on the random date IDRSA model (Table 4 and Supplementary Table 3), we can see the number of individuals aged ≥50 years expected to get CDI by risk score threshold or range demonstrated by the following examples. (1) Risk scores ≥0.05 will target 0.07% of individuals (707 of 972 172) and 1.6% of all CDI cases (30 of 1918) (Table 4). (2) Alternatively, lower risk scores of ≥0.01 will target 2.1% of a random population aged ≥50 years (20 676 of 972 172) and 18.8% of all CDI cases (360 of 1918) (Table 4), which may be more feasible for enrollment of nonhospitalized individuals in *C difficile* vaccine trials. (3) In the 0.05 to 0.10 risk score range, we predict 4.2% will get CDI (Supplementary Table 3).

Supplemental analyses limited to 1–30 days after the random IDRSA model revealed that of the 157 CDI cases that occurred during this period, only 1 had a risk score threshold of developing CDI in the first 30 days above 0.05%, indicating that our model was not overly affected by excluding cases during the 1–30 days after a random date.

The potential risk factors included in each risk score model, along with coefficients generated from logistic regression (Supplementary Tables 4 and 5), can be used to calculate a risk score for individuals. For example, for each of the 19 risk factors in the simplified model noted in Supplementary Table 5, individuals would be assigned "1's" for present risk factors and "0's" for absent risk factors. Each "1" and "0" would be multiplied by the coefficient provided. All coefficients (including the baseline intercept) would then be added together and the sum exponentiated to yield the probability that an individual with this risk score recruited for a clinical trial would have CDI in 31–365 days following a random date.

DISCUSSION

In this large study consisting of approximately 2.2 million KPNC members and 9986 CDI cases, we identified CDI risk factors, which we used to estimate the risk of developing CDI 31 to 365 days after either a hospital discharge or a random date. The hospital discharge IDRSA model was excellent in predicting future CDI (C-statistic 0.848), whereas the random IDRSA model predicted CDI reasonably well (C-statistic 0.722). Although this study newly identified that having at least 10 outpatient visits in the past year was a risk factor for CDI (53% in CDI cases vs 16% in the comparison), most factors in our

Table 3. Validation Model Risk Scores for Developing a *Clostridioides difficile* Infection 31 to 365 Days After a Hospital Discharge Index Date for Risk Score Assessment for Individuals Aged 18+ Years Using Various Risk Score Thresholds, Kaiser Permanente Northern California, May 2011 to July 2014

Risk Score Threshold	Total Number of Eligible Hospital Discharges (% of All Hospital Discharges)	Number Who Had CDI in 31–365 Days After Hospital Discharge IDRSA	Percentage of Hospital Discharges in This Row Who Had CDI (Col 3/Col 2)	Percentage of Total CDI Cases (n = 3423)
Risk ≥0.00	104 518 (100)	3423	3.3%	100%
Risk ≥0.005	83 989 (80.4)	3372	4.0%	98.5%
Risk ≥0.01	54 248 (51.9)	3188	5.9%	93.1%
Risk ≥0.02	31 403 (30)	2833	9.0%	82.8%
Risk ≥0.03	21 328 (20.4)	2491	11.7%	72.8%
Risk ≥0.04	15 825 (15.1)	2197	13.9%	64.2%
Risk ≥0.05	12 423 (11.9)	1957	15.8%	57.2%
Risk ≥0.10	5486 (5.2)	1205	22.0%	35.2%
Risk ≥0.15	3203 (3.1)	844	26.4%	24.7%
Risk ≥0.20	2129 (2)	647	30.4%	18.9%
Risk ≥0.25	1514 (1.4)	483	31.9%	14.1%
Risk ≥0.30	1146 (1.1)	379	33.1%	11.1%

Abbreviations: CDI, *Clostridioides difficile* infection; Col, column; IDRSA, Index Date for Risk Score Assessment.

Table 4. Validation Model Risk Scores for Developing a *Clostridioides difficile* Infection 31 to 365 Days After a Random Index Date for Risk Score Assessment (IDRSA) for Individuals Aged 50+ Years Using Various Thresholds, Kaiser Permanente Northern California May 2011 to July 2014

Risk Score Threshold	Total Number of Eligible Individuals N = 972 172 (% of total)	Number Who Had CDI in 31–365 days After Random IDRSA	Percent of Total Population in This Row Who Had CDI (Col 3/Col 2)	Percent of Total CDI Cases (n = 1918)
Risk score ≥0.00	972 172	1918	0.2%	100%
Risk ≥0.005	55 726 (5.7)	638	1.1%	33.3%
Risk ≥0.01	20 676 (2.1)	360	1.7%	18.8%
Risk ≥0.02	6550 (0.7)	167	2.6%	8.7%
Risk ≥0.03	2870 (0.3)	98	3.4%	5.1%
Risk ≥0.04	1352 (0.1)	51	3.8%	2.7%
Risk score ≥0.05	707 (0.07)	30	4.2%	1.6%
Risk score ≥0.10	37 (0.004)	2	5.4%	0.1%
Risk score ≥0.11	23 (0.002)	2	8.7%	0.1%
Risk score ≥0.12	8 (0.0008)	1	12.5%	0.1%
Risk score ≥0.13	3 (0.0003)	0	0.0%	0.0%

Abbreviations: CDI, *Clostridioides difficile* infection; Col, column; IDRSA, Index Date for Risk Score Assessment.

study, such as older age, hospitalization (acute and long-term care), ED visits, outpatient visits, antibiotic use, proton pump inhibitor use, and specific comorbidities, have been reported previously [1, 3, 10, 16, 20], which confirms the importance of these CDI risk factors and lends credibility to our prediction models. Overall, our model was able to identify persons who, at the time of discharge from the hospital, were at high risk for developing CDI during the subsequent year.

Our risk score can be interpreted as the probability of developing CDI during the follow-up interval, in a practice setting similar to KPNC during 2011–2014. A risk score that easily identifies individuals at high risk for CDI would be useful for *C difficile* vaccine trials. The ability of such a tool to enrich the study population for individuals more likely to have CDI within the next year would allow a vaccine efficacy study to reach the required endpoints for conclusion in a meaningful timeframe with a feasible sample size. These tools could also be used to target individuals at high risk for CDI who may benefit from precautionary measures such as avoidance of unnecessary antibiotics and proton-pump inhibitors and isolation procedures for infection control.

Other risk score models for CDI have been developed, but most were limited in scope by focusing on only hospital-acquired and/or institution-specific CDI [9, 11–15, 17] and/or had small samples (eg, <300 CDI cases vs ~10 000 cases in our study) [9, 11, 13, 17]. Kuntz et al [10] predicted risk of CDI after outpatient visits, but their study was also limited in size (620 CDI cases). A study comparable to ours by Zilberberg et al [16] collected risk factors from a large Medicare sample to develop a simplified model that assigned risk based on 22 predictors. Their model also performed well in predicting future CDI (C-statistic 0.864). Although their study was more geographically diverse than ours (US Medicare claims vs Northern California), our study included a larger age range (≥18 years in our study vs ≥65) and a more racially/ethnically diverse population (70% white in our study vs 90%).

Study strengths include KPNC's large membership size, which increased power and confidence in the data, and our ability to access the complete medical record for all individuals in the study. Kaiser Permanente Northern California's integrated EMR allowed us to both identify all individuals with relevant risk factors and to follow them over time for subsequent development of CDI. We also created a simplified model using a random date that could be applicable to different healthcare settings.

This study has limitations. First, CDI risk at KPNC may differ from elsewhere due to different hospital and geographic settings. Second, the usefulness of our findings may be limited in settings where extensive data are unavailable. Third, although KPNC's CDI testing procedure implemented in 2011 was in use during the study period, data from 2010 used to confirm no prior CDI may have been misclassified due to different testing procedures used earlier. Fourth, our ability to capture complete data for KPNC members in long-term care facilities outside of KPNC was limited and may have missed CDI if no CDI care was received at KPNC. Fifth, these are predictive models rather than causal models, and important causal factors may be omitted arbitrarily by our variable selection processes. The importance of a variable in our predictive models may not reflect its importance in causing CDI. Sixth, our validation dataset was selected randomly rather than temporally so we did not examine how a predictive algorithm developed from earlier data predicts CDI in a later time period. Seventh, we did not make use of newer machine learning methods for variable selection and cross-validation (to address overfitting). Finally, although previous CDI is a well known risk factor for recurrence [1, 21], our study focused on incident CDI and not recurrent CDI.

One goal of our study was to create the best possible risk score model for predicting future CDI that could be used to plan for *C difficile* vaccine clinical efficacy trials. Our complex model, which incorporates many risk factors, demonstrates that we could identify 57.2% of hospitalized individuals expected to get CDI by focusing on just hospitalized individuals with risk scores above 0.05 (approximately 12% of all hospitalizations). Creating similar risk scores based on many risk factors may be practical in other healthcare settings that also have extensive EMRs. A secondary goal was to create a simplified risk score model that can be scored with readily available risk factors in a wide range of clinical settings without extensive EMRs. We also expanded the population for this model to include individuals in the outpatient setting because the incidence of CDI in this population has been increasing. Although a simplified model such as the random IDRSA model sacrifices predictive power and does not predict future CDI with the same accuracy as the hospital discharge IDRSA model, this simplified version more easily identifies high-risk individuals based on the most important risk factors. Further, this simplified risk score model has the advantage of greater flexibility in situations such as recruiting for a clinical trial (for example, sending a recruitment letter to potential subjects before a planned visit).

CONCLUSIONS

In conclusion, we have confirmed important risk factors of CDI. We used these and other results to inform modeling and create 2 types of predictive risk scores—one focused on a hospital setting and one that includes the outpatient/community setting. These models can be used to plan for and recruit subjects into future *C difficile* vaccine clinical efficacy trials.

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Acknowledgments

Authors' contributions. All authors were involved in the conception of the design of the study and participated in the development of this manuscript: L. A. wrote the first draft; J. T., E. L., and J. H. collected or generated the data; N. P. K., E. L., B. F., J. T., J. H., and L. A. analyzed study data; all authors interpreted study data. All authors revised the manuscript critically for important intellectual content. All authors approved the final version before submission. All authors agree to be accountable for all aspects of the work. The corresponding author had final responsibility to submit for publication.

Financial support. Pfizer, Inc. funded this study and was involved in all aspects of the study, including study design and interpretation of the data.

Potential conflicts of interest. H. Y., B. C., E. G., and J. L. are employees of Pfizer, Inc., the funder of the study. The study was conducted and analyzed at Kaiser Permanente, where all the other authors are employed. N. K. has received research grant support for unrelated studies from GlaxoSmithKline, Sanofi Pasteur, Pfizer, Merck, MedImmune (now AstraZeneca), Protein Sciences (now Sanofi Pasteur), and Dynavax. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Lessa FC, Mu Y, Bamberg WM, et al. Burden of *Clostridium difficile* infection in the United States. N Engl J Med **2015**; 372:825–34.
2. Steiner C, Barrett M, Sun Y. HCUP projections: *Clostridium difficile* hospitalizations 2004 to 2015. HCUP Projections Report #2015-02. Available at: https://www.hcup-us.ahrq.gov/reports/projections/2015-02.pdf. Accessed 5 December 2018.
3. Baxter R, Ray GT, Fireman BH. Case-control study of antibiotic use and subsequent *Clostridium difficile*-associated diarrhea in hospitalized patients. Infect Control Hosp Epidemiol **2008**; 29:44–50.
4. Knight CL, Surawicz CM. *Clostridium difficile* infection. Med Clin North Am **2013**; 97:523–36, ix.
5. Bradshaw W, Bruxelle JF, Kovacs-Simon A, et al. Molecular features of lipoprotein CD0873: a potential vaccine against the human pathogen *Clostridioides difficile*. J Biol Chem **2019**; 294:15850.
6. Senoh M, Iwaki M, Yamamoto A, et al. Development of vaccine for *Clostridium difficile* infection using membrane fraction of nontoxigenic *Clostridium difficile*. Microb Pathog **2018**; 123:42–6.
7. ClinicalTrials.gov Identifier: NCT03090191 - A Phase 3, Placebo-controlled, randomized, observer-blinded study to evaluate the efficacy, safety, and tolerability of a *Clostridium difficile* vaccine in adults 50 years of age and older (Pfizer Protocol No. B5091007). Available at: https://clinicaltrials.gov/ct2/show/NCT03090191. Accessed 9 December 2018.
8. ClinicalTrials.gov Identifier: NCT01887912 - Efficacy, Immunogenicity, and Safety Study of *Clostridium difficile* toxoid vaccine in subjects at risk for *C. difficile* infection (Cdiffense™) (Sanofi Pasteur Protocol No. H-030-014). Available at: https://clinicaltrials.gov/ct2/show/NCT01887912. Accessed 9 December 2018.
9. Chandra S, Thapa R, Marur S, Jani N. Validation of a clinical prediction scale for hospital-onset *Clostridium difficile* infection. J Clin Gastroenterol **2014**; 48:419–22.
10. 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**; 21:256–62.
11. Na X, Martin AJ, Sethi S, et al. A multi-center prospective derivation and validation of a clinical prediction tool for severe *Clostridium difficile* infection. PLoS One **2015**; 10:e0123405.
12. Oh J, Makar M, Fusco C, et al. A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. Infect Control Hosp Epidemiol **2018**; 39:425–33.
13. Press A, Ku B, McCullagh L, et al. Developing a clinical prediction rule for first hospital-onset *Clostridium difficile* infections: a retrospective observational study. Infect Control Hosp Epidemiol **2016**; 37:896–900.
14. Wiens J, Campbell WN, Franklin ES, et al. Learning data-driven patient risk stratification models for *Clostridium difficile*. Open Forum Infect Dis **2014**; 1:ofu045.
15. Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inform Assoc **2014**; 21:699–706.
16. Zilberberg MD, Shorr AF, Wang L, et al. Development and validation of a risk score for *Clostridium difficile* infection in Medicare beneficiaries: a population-based cohort study. J Am Geriatr Soc **2016**; 64:1690–5.
17. Baggs J, Yousey-Hindes K, Ashley ED, et al. Identification of population at risk for future *Clostridium difficile* infection following hospital discharge to be targeted for vaccine trials. Vaccine **2015**; 33:6241–9.
18. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol **1992**; 45:613–9.
19. SAS. Usage Note 39724: ROC analysis using validation data and cross validation. Available at: http://support.sas.com/kb/39/724.html. Accessed 5 December 2018.
20. Guh AY, Adkins SH, Li Q, et al. Risk factors for community-associated *Clostridium difficile* infection in adults: a case-control study. Open Forum Infect Dis **2017**; 4:ofx171.
21. Eyre DW, Walker AS, Wyllie D, et al; Infections in Oxfordshire Research Database. Predictors of first recurrence of *Clostridium difficile* infection: implications for initial management. Clin Infect Dis **2012**; 55 (Suppl 2):S77–87.