ORIGINAL ARTICLE

Estimating Local Costs Associated With *Clostridium difficile* Infection Using Machine Learning and Electronic Medical Records

Theodore R. Pak, PhD;[1] Kieran I. Chacko, BA, BSc;[1] Timothy O'Donnell, BS;[1] Shirish S. Huprikar, MD;[2] Harm van Bakel, PhD;[1] Andrew Kasarskis, PhD;[1] Erick R. Scott, MD, MHS[1]

BACKGROUND.   Reported per-patient costs of *Clostridium difficile* infection (CDI) vary by 2 orders of magnitude among different hospitals, implying that infection control officers need precise, local analyses to guide rational decision making between interventions.

OBJECTIVE.   We sought to comprehensively estimate changes in length of stay (LOS) attributable to CDI at a single urban tertiary-care facility using only data automatically extractable from the electronic medical record (EMR).

METHODS.   We performed a retrospective cohort study of 171,938 visits spanning a 7-year period. In total, 23,968 variables were extracted from EMR data recorded within 24 hours of admission to train elastic-net regularized logistic regression models for propensity score matching. To address time-dependent bias (reverse causation), we separately stratified comparisons by time of infection, and we fit multistate models.

RESULTS.   The estimated difference in median LOS for propensity-matched cohorts varied from 3.1 days (95% CI, 2.2–3.9) to 10.1 days (95% CI, 7.3–12.2) depending on the case definition; however, dependency of the estimate on time to infection was observed. Stratification by time to first positive toxin assay, excluding probable community-acquired infections, showed a minimum excess LOS of 3.1 days (95% CI, 1.7–4.4). Under the same case definition, the multistate model averaged an excess LOS of 3.3 days (95% CI, 2.6–4.0).

CONCLUSIONS.   In this study, 2 independent time-to-infection adjusted methods converged on similar excess LOS estimates. Changes in LOS can be extrapolated to marginal dollar costs by multiplying by average costs of an inpatient day. Infection control officers can leverage automatically extractable EMR data to estimate costs of CDI at their own institutions.

Clostridium difficile infection (CDI) is the most frequently reported healthcare-associated infection (HAI) in the United States[1] and the major infective cause of nosocomial diarrhea in developed countries,[2] incurring billions of dollars in excess medical costs per year.[3] Estimates of the per-patient cost of CDI have varied from $2,871 to $122,318 due to differences in methodology, patient inclusion criteria, and regional costs.[4–6] Given the high hospital-to-hospital variability of these costs,[7,8] infection control officers, hospital administrators, and clinicians would benefit from estimates tailored to their particular populations and healthcare practices. Concretely defining the potential economic savings of CDI prevention would empower stakeholders to prudently choose among the many available validated interventions.[9,10]

Measuring costs within healthcare systems is notoriously difficult; many hospitals do not have access to itemized reimbursement data linked to medical records.[11] Even institutions that have informatics retrospectively linking these data have relied on the curation of select variables and chart review to estimate attributable CDI cost.[12–14] Nevertheless, electronic medical record (EMR) systems are used by most first-world acute-care facilities.[15,16] Part of the rationale for these systems is that hospitals may leverage EMR data for optimal decision making by inferring causal relationships from raw observations during routine care.[17–19] An analysis based on automatically extractable data from an EMR that quantifies preventable hospital costs, such as those attributable to an HAI like CDI, would be of great value in building a continuously learning healthcare system.[20] EMRs contain many structured fields relevant to this analysis, including diagnosis codes and lab results demonstrating onset of HAIs; thousands of variables for procedures, problems, and medications that can serve as covariates for adjustment in observational studies; and importantly, the length of stay (LOS) for each visit, which is the primary contributor to excess costs for most HAIs, including CDI.[3,21,22]

The goal of this study was to generate a robust estimate of local cost associated with CDI using data that are automatically

extractable from a typical EMR. We used all available structured data recorded within 24 hours of admission in the EMR (including >20,000 variables, such as medications reported and administered, abnormal lab values, and problem list entries) to build fully data-driven models for CDI risk using a machine-learning algorithm to avoid the potential bias of preselected covariates and manual chart review. CDI risk models trained on uncurated data from EMRs have already outperformed models that only incorporate variables for known risk factors, indicating that CDI risk may be nuanced in particular care settings.[23] We then use these trained CDI risk models for propensity score matching, which allowed estimation of changes in LOS associated with CDI. Most previous studies of CDI cost have not accounted for the possibility that longer LOS increases the risk of CDI (ie, reverse causation), and therefore likely overestimate the cost of CDI.[7,24] To adjust for this, we stratified our analysis by the time of CDI diagnosis to find the change in LOS conditional on minimal prior exposure to the hospital environment. Finally, we compared these results to a multistate model of competing time-dependent risks between discharge and the onset of CDI.

METHODS

Data Source

This study was conducted at The Mount Sinai Hospital, a 1,171-bed tertiary-care hospital in New York City. Records of warehoused adult inpatient EMR visit data were deidentified using the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor method, 45 CFR §164.514(b)(2). Data were collected on demographics, LOS, time of death, admission sources, reported medications, and the presence of a "008.45" *International Classification of Disease, Ninth Revision* (ICD-9) principal or secondary visit diagnosis code denoting "Intestinal infection due to *Clostridium difficile*." Furthermore, all records of medications administered, abnormal lab results, surgery procedure codes, or problem list ICD-9 codes within the first 24 hours after admission were collected as Boolean variables (ie, presence or absence). All variables that were uniform across the study population were dropped from the dataset. The relationships between collected data elements are summarized in Figure 1A. The Mount Sinai Institutional Review Board deemed this research to be exempt from the need for approval.

Study Population

The cohort included all patients 18 years of age or older admitted between January 1, 2009, and October 22, 2015 (Figure 1B). For each patient, visits following the first recorded visit in the time range were excluded so that each patient corresponded to a single visit. Visits involving a patient death, defined as a recorded time of death within 24 hours after discharge, were excluded (2,682 adult patients; 1.5%). Visits with missing or invalid date information were excluded (<0.01% of all records).

Study Design

Prior studies vary on the use of ICD-9 discharge codes versus positive laboratory tests to define CDI cases[5,6] and identify differing positive predictive values for immunoassay and nucleic acid–based laboratory tests.[25–27] To ensure maximally robust results and to allow comparison with prior studies, we repeated our analysis for 5 definitions of CDI:

Definition 1: An "008.45" ICD-9 visit diagnosis code
Definition 2: ≥1 positive stool toxin enzyme immunoassay (EIA) lab result
Definition 3: ≥1 positive stool toxin polymerase chain reaction (PCR) lab result
Definition 4: Definition 2 or definition 3
Definition 5: Definition 1, 2, or 3

Our study period included both a period during which the EIA assay was the standard hospital laboratory test (~3 years) followed by a period during which the PCR assay was standard (~4 years). For case cohorts involving definitions 2 and 3, comparisons were only permitted with controls from the period during which that same test was standard. The hospital laboratory protocol requires unformed stool samples for either toxin assay.

Statistical Analysis

Details of propensity model development, matching, evaluation of matching performance, and LOS comparisons are available in Supplementary Methods. Briefly, propensity models for CDI based on the 5 case definitions were trained using logistic regression with elastic net regularization. After exact matching on gender and age bins, nearest-neighbor 1:1 matching on the propensity score was performed with a caliper of 0.2 standard deviations of the logit of the propensity score (Figure S1).[28] Matching was repeated using the matched controls against remaining unmatched controls to create a rematched cohort, testing whether matching alone is associated with changes in LOS. For each case definition of CDI, differences of the median LOS between cases and matched controls were calculated, and statistical significance was determined using with the 2-sided Mann-Whitney U test. Although violation of the proportional hazards assumption (Figure S2) pre-empted traditional Cox survival analysis, nonparametric Kaplan-Meier estimates of the time-dependent risk of discharge were plotted for matched cohorts.

To further address the possible effect of time to infection on CDI risk and measured LOS differences, we repeated the analysis for definition 4, stratifying by the time of the first positive toxin assay using 3 ranges: 0–3 days, 3–8 days, and ≥8 days. Propensity models were again fitted to each of these case cohorts for matching as described previously, with the added condition that controls discharged before the start of the CDI time window were ineligible for matching.[29] LOS comparisons followed the same procedure as above. Furthermore, we fit a nonparametric multistate model consistent with previous studies,[7,24,30] under which the mean excess LOS was estimated as the average difference in LOS between patients that had or
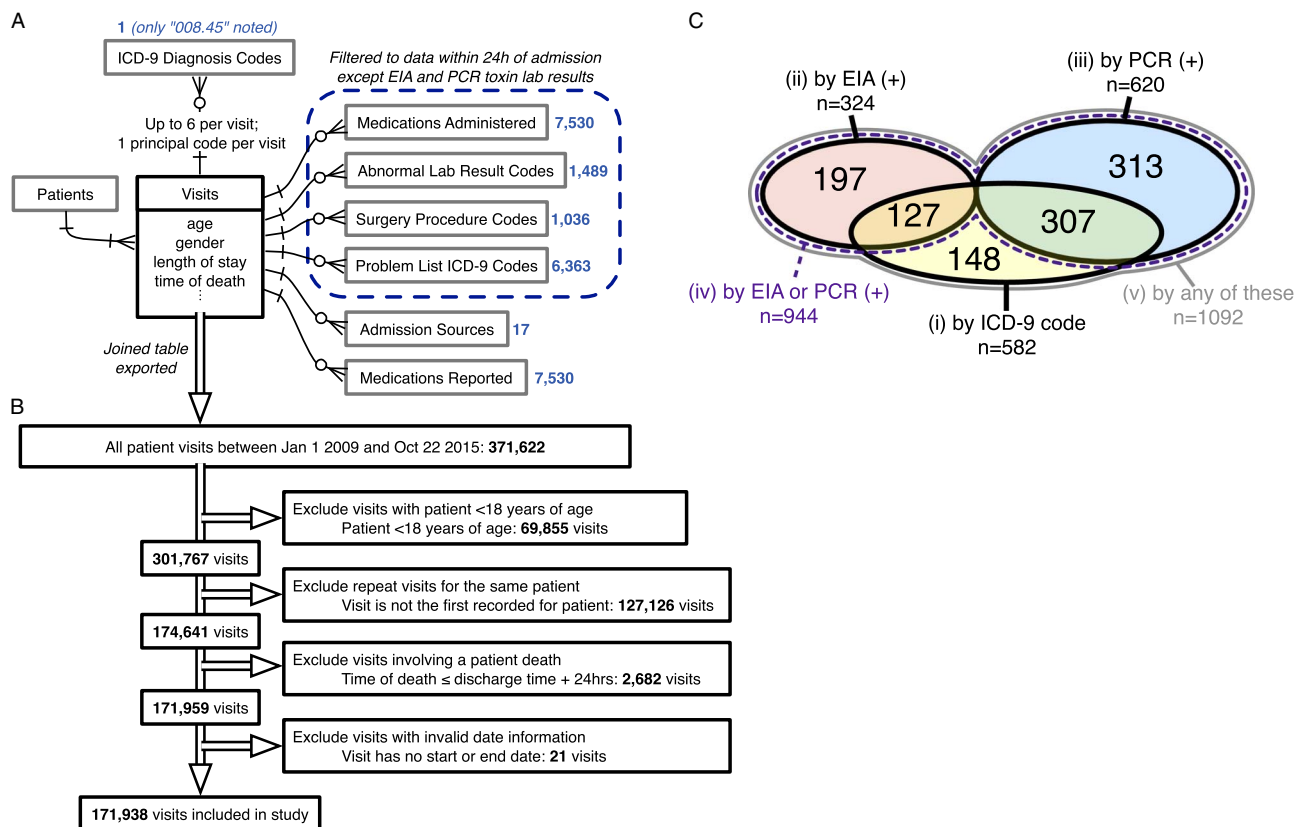
FIGURE 1. Data sources, inclusion/exclusion criteria, and cohort sizes before matching. (A) Entity-relationship diagram for all EMR data used to generate models of CDI propensity, using information engineering notation.[44] Boxes represent tables of entities with any directly associated attributes (fields) listed below; single lines represent relationships, with arrowheads indicating the cardinality of each side of the relationship; crow's foot arrowhead with circle represents "zero or more"; crow's foot arrowhead with a cross stroke represents "1 or more"; cross-stroke arrowhead represents "exactly one." Blue numbers indicate the number of variables extracted from each associated table for each visit. (B) Inclusion/exclusion procedure for the present study. Double-line arrows indicate the procession of visit records. (C) Venn diagram of case cohort sizes for each of the 5 CDI case definitions before matching, including sizes of all intersections between case definitions (overlaps). Areas are not to scale. There is no intersection between definitions 2 and 3 because only the first positive toxin assay result for each visit was examined. Definition 4, "by EIA or PCR (+)," is a strict superset of definitions 2 and 3. Definition 5, "by any of these," is a strict superset of definitions 1, 2, and 3. Sizes of matched case cohorts are provided in Table 1. EMR, electronic medical record; CDI, *Clostridium difficile* infection.

had not transitioned through the infected state for all timepoints, weighted by the distribution of times spent in the uninfected state. Analyses were performed in R 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria); all software code is available at https://github.com/powerpak/cdi-cost.

RESULTS

In total, 371,622 records of visits during the study time range were queried from the EMR, with 23,968 variables extracted for each visit (Figure 1A and 1B). After filtering for the index visit per adult patient and excluding deaths and invalid dates, 171,938 visits were deemed eligible for inclusion and were classified into 5 overlapping case definitions for CDI. Case cohort sizes before matching and their overlaps are depicted in Figure 1C. Regularized logistic regression models predicting

the risk of CDI acquisition were fitted to EMR data from the first 24 hours of each admission for each case definition, with consistently high predictive performance (Supplementary Methods; Figure S3).

For each case definition, >75% of cases were successfully matched by propensity score to controls (Figure 1C and Table 1). The groups are well matched on demographics and propensity scores (Table 1 and Figure S4). Differences in the median LOS between matched case and control cohorts for all CDI case definitions were strongly statistically significant, although the magnitude of the differences varied greatly between definitions (Figure 2A). The differences in the median LOS, by case definition, were definition 1 (by ICD-9 code), 3.1 days (95% confidence interval [CI], 2.2–3.9); definition 2 (by positive toxin EIA), 10.1 days (95% CI, 7.3–12.2); definition 3 (by positive toxin PCR), 6.6 days (95% CI, 5.0–8.1);

TABLE 1.   Demographic Characteristics of the Study Population and Matched Cohorts

		Matched Cohorts for Each CDI Case Definition					
		Definition 1: By ICD-9 Code			Definition 2: By EIA (+)		
	No. (%)	No. (%)			No. (%)		
Characteristic	All Patients (n = 171,936)	All Controls (n = 171,356)	Matched Cases & Controls[a] (n = 489)	SMD After Matching (P Value)	All Controls (n = 73 647)	Matched Cases & Controls[a] (n = 274)	SMD After Matching (P Value)
Female sex	101,964 (59)	101,638 (59)	278 (57)	0 (1)	44,132 (60)	145 (53)	0 (1)
Age[b]				.016 (0.86)			.018 (0.79)
18–29	22,344 (13)	22,266 (13)	69 (14)		9,552 (13)	22 (8)	
30–44	39,003 (23)	38,898 (23)	86 (18)		16,451 (22)	26 (9)	
45–59	37,234 (22)	37,129 (22)	90 (18)		15,956 (22)	58 (21)	
60–74	43,946 (26)	43,802 (26)	122 (25)		18,407 (25)	83 (30)	
75–90	26,167 (15)	26,041 (15)	106 (22)		11,817 (16)	70 (26)	
≥90	3,244 (2)	3,220 (2)	16 (3)		1,464 (2)	15 (5)	

	Matched Cohorts for Each CDI Case Definition (Extended)					
	Definition 3: By PCR (+)			Definition 4: By EIA or PCR (+)		
	No. (%)			No. (%)		
Characteristic	All Controls (n = 97,351)	Matched Cases & Controls[a] (n = 493)	SMD After Matching (P Value)	All Controls (n = 170 994)	Matched Cases & Controls[a] (n = 788)	SMD After Matching (P Value)
Female sex	57,340 (59)	254 (52)	0 (1)	101,469 (59)	408 (52)	0 (1)
Age[b]			.005 (0.99)			.003 (0.98)
18–29	12,714 (13)	47 (10)		22,265 (13)	79 (9)	
30–44	22,430 (23)	72 (15)		38,879 (23)	124 (13)	
45–59	21,069 (22)	117 (24)		37,025 (22)	209 (23)	
60–74	25,273 (26)	136 (28)		43,680 (26)	266 (29)	
75–90	14,120 (15)	114 (23)		25,936 (15)	231 (24)	
≥90	1,745 (2)	7 (1)		3,209 (2)	35 (3)	

	Matched Cohorts for Each CDI Case Definition (Extended)		
	Definition 5: By Any of the Criteria		
	No. (%)		
Characteristic	All Controls (n = 170,846)	Matched Cases & Controls[a] (n = 945)	SMD After Matching (P Value)
Female sex	101,390 (59)	493 (52)	0 (1)
Age[b]			.004 (0.93)
18–29	22,245 (13)	87 (9)	
30–44	38,845 (23)	134 (14)	
45–59	36,999 (22)	208 (22)	
60–74	43,643 (26)	267 (28)	
75–90	25,912 (15)	217 (23)	
≥90	3,202 (2)	32 (3)	

NOTE. CDI, *Clostridium difficile* infection; ICD-9, *International Classification of Diseases Ninth Revision*; EIA, enzyme immunoassay; PCR, polymerase chain reaction; SMD, standardized mean difference.
[a]Separate columns are unnecessary because 1:1 exact matching was performed on the characteristics shown, and therefore all values are identical.
[b]SMD is shown for age treated as a continuous variable; coarsened exact matching was performed using the listed age ranges.
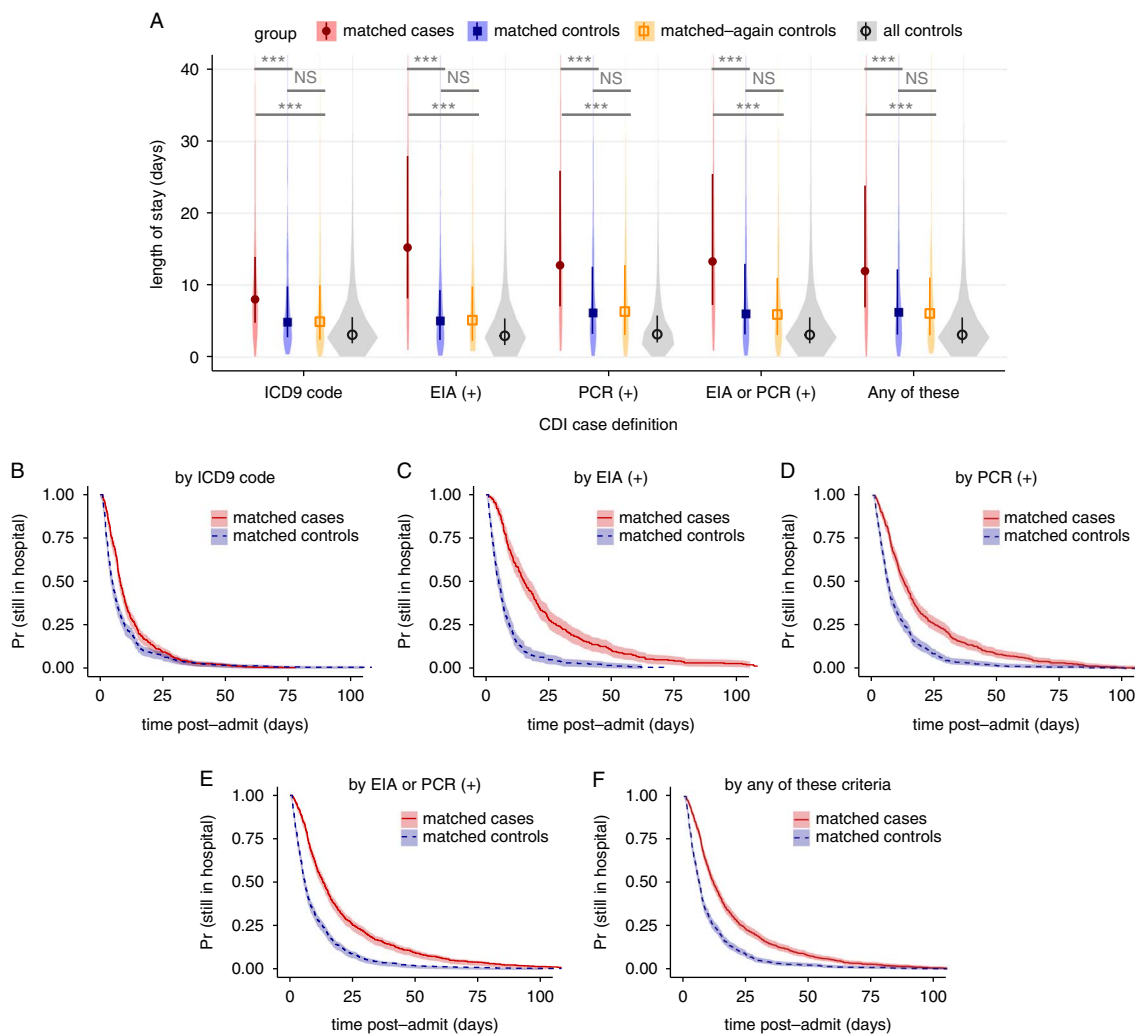
FIGURE 2. Changes in length of stay for 5 case definitions of *Clostridium difficile* infection, not accounting for time of infection. (A) Violin plots of the distributions in length of stay for matched cases, matched controls, matched-again controls, and all controls, for each of the 5 case definitions. Darker points and vertical bars depict the median and interquartile range for each group. Horizontal bars depict Mann-Whitney *U* tests for significance of differences between groups (***, Bonferroni-corrected $P < .001$; NS, not significant [$P > .1$]). (B–F) Kaplan-Meier plots of the time-dependent probability for a patient to still be in the hospital, comparing matched cases and controls for each case definition of CDI. Shaded areas depict 95% confidence intervals calculated from standard errors. CDI, *Clostridium difficile* infection; ICD-9, *International Classification of Diseases Ninth Revision*; EIA, enzyme immunoassay; PCR, polymerase chain reaction.

definition 4 (by either toxin assay), 7.2 days (95% CI, 5.8–8.3); and definition 5 (by any of these), 5.7 days (95% CI, 4.5–6.6). There were no significant differences in LOS for a second round of matching between matched controls and remaining controls (rematched controls) for any of the case definitions (Figure 2A). Kaplan-Meier curves for the time-dependent risk of being discharged from the hospital showed significant differences between matched case and control cohorts up to post-admission day 60 for all case definitions except ICD-9 code (Figure 2B–F).

Estimates of LOS associated with CDI are inflated by dependencies on time-to-infection; longer preinfection LOS increases CDI risk (ie, reverse causation) and leads to overestimates

in attributable cost.[7,24] Therefore, we performed 2 follow-up analyses to account for this. First, we stratified the LOS comparison by the time of CDI diagnosis for case definition 4 into case cohorts of 0–3 days, 3–8 days, and ≥8-days, training new propensity models for rematching, with similar performance (Figure S5). Because 3 days is a typical cutoff for differentiating community-acquired (CA) from healthcare-associated (HA) CDI,[25,31] these strata were named "CA," "early HA," and "late HA," respectively. As suspected, stratification revealed a positive correlation between time of diagnosis and CDI-associated difference in LOS (Figure 3A). The differences in medians were (1) for CA, 2.5 days (95% CI, 1.2–3.4); (2) for early HA, 3.1 days (95% CI, 1.8–4.4); and (3) for late HA, 14.0 days
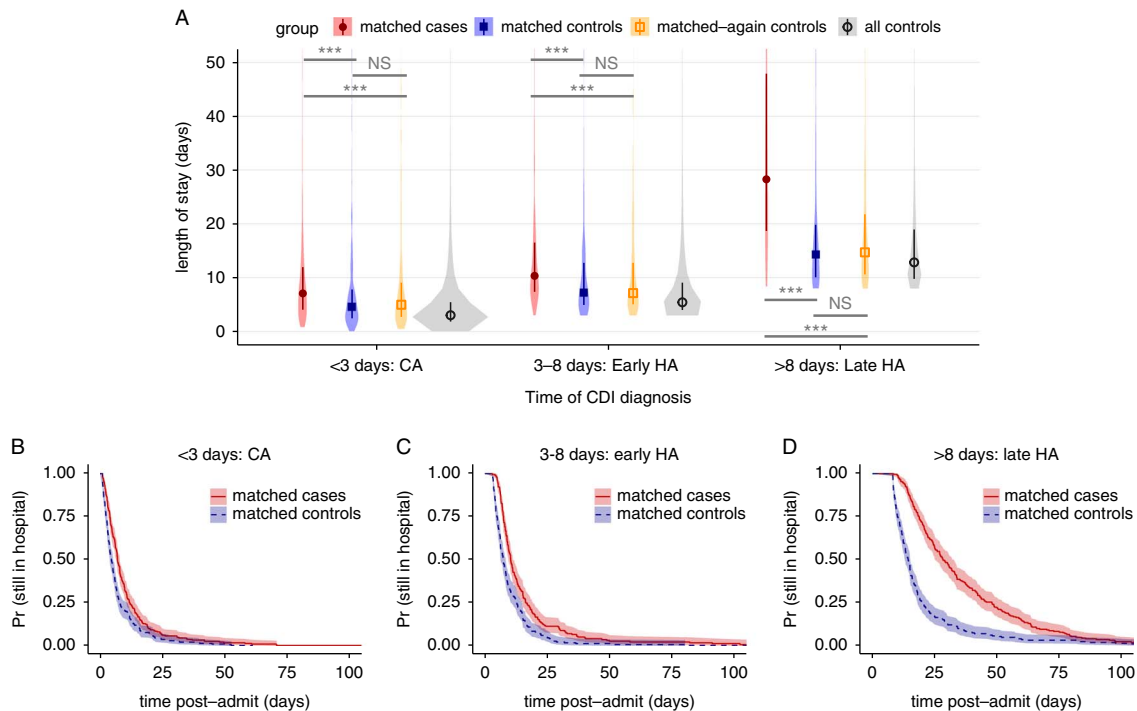
FIGURE 3. Changes in length of stay for *Clostridium difficile* infection defined by any positive toxin assay, stratified by the time to infection. (A) Violin plots of the distributions in length of stay for matched cases, matched controls, rematched controls, and all controls, for 3 ranges of the result time for the first positive toxin assay. Points and vertical bars depict the median and interquartile range for each group. Horizontal bars depict Mann-Whitney U tests for significance of differences between groups (***, Bonferroni-corrected $P < .001$; NS, not significant [$P > .1$]). (B–D), Kaplan-Meier plots of the time-dependent probability for a patient to still be in the hospital, comparing matched cases and controls for the same 3 ranges of the time of the first positive toxin assay. Shaded areas depict 95% confidence intervals calculated from standard errors. CDI, *Clostridium difficile* infection; CA, community acquired; HA, healthcare associated.

(95% CI, 9.9–17.1). All comparisons between matched cases and controls were again strongly statistically significant, and comparisons with rematched controls were not significant (Figure 3A). Kaplan-Meier plots likewise confirmed a correlation between time of CDI diagnosis and differences in time-dependent discharge risk (Figure 3B–D).

To further address reverse causation, we fit a multistate model similar to previously published studies[7,24,30] that explicitly estimates time-dependent, competing risks of transitioning to CDI versus discharge. Figure 4A depicts the model's states and transitions. After fitting the model for the case definitions with a time of diagnosis (definitions 2, 3, and 4), the expected remaining LOS can be compared across cohorts that have already transitioned to the CDI infected state versus those that are still CDI negative at any given timepoint (Figure 4B–D). To summarize the overall relationship between CDI and LOS, differences in LOS were weighted by the distribution of times spent in the initial state and averaged. The average differences for each case definition were: definition 2 (by positive toxin EIA), 3.0 days (95% CI, 2.0–4.0); definition 3 (by positive toxin PCR), 3.5 days (95% CI, 2.7–4.5); and definition 4 (by either toxin assay), 3.3 days (95% CI, 2.6–4.0). Notably, the 95% CI for the difference in the definition 4

cohort overlaps the 3.1-day difference for the "early HA" stratum of the propensity-matched analysis in the same cohort.

DISCUSSION

This study examined nearly 7 years of uncurated EMR data for a single hospital and determined associated costs of CDI as defined by either visit diagnosis codes or lab results. In the analysis unadjusted for time to infection, differences in LOS were often greater than national averages from similar unadjusted studies,[3,5,6] but changes in the case definition resulted in substantial changes in the estimated differences in LOS. Although 2 hospitals reported good concordance between ICD-9 codes and CDI toxin assay results,[32,33] this is not necessarily the case for all hospitals. We found that 75% of ICD-9 coded visits involved a positive toxin assay, while only 46% of visits with a positive toxin assay had the ICD-9 code (Figure 1C). Changes in LOS were not significantly different between EIA and PCR toxin assays, although our study was limited by a smaller sample size for EIA-positive cases. Toxin assays are likely a more reliable CDI definition given their basis in clinical symptoms and evidence for CDI, whereas medical coding suffers from biases introduced by billing and reimbursement.[34,35]
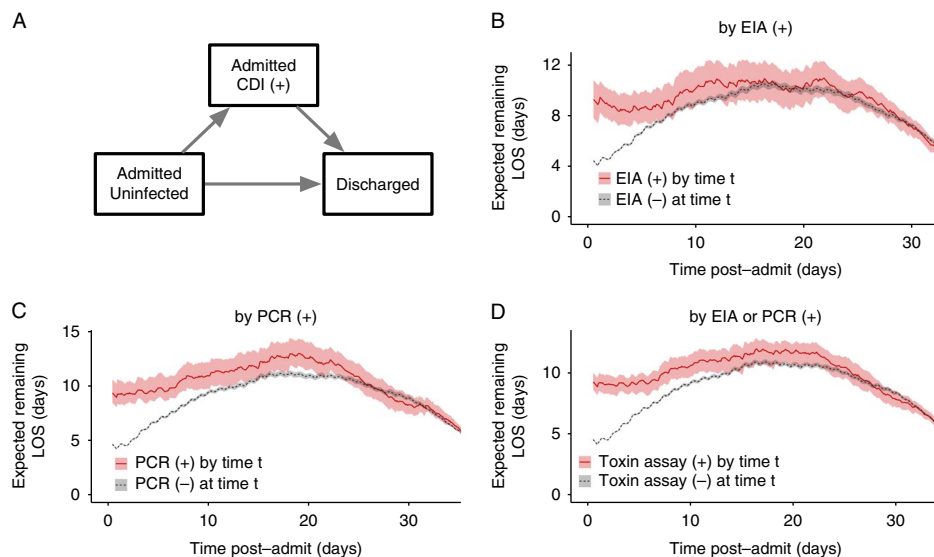
FIGURE 4. Multistate model of expected remaining length of stay for *Clostridium difficile* infection case definitions involving toxin assays. (A) The 3 states of the multistate model and allowed transitions. Patients may only transition in the direction of the arrows. (B–D) Expected remaining LOS for each post-admission time *t* depending on whether the patient has had a positive (+) toxin assay by that timepoint, for each of the case definitions involving toxin assays. Shaded areas depict 95% confidence intervals calculated from 1,000 bootstrap samples. CDI, *Clostridium difficile* infection; EIA, enzyme immunoassay; PCR, polymerase chain reaction; LOS, length of stay.

Treating CDI as a baseline condition by ignoring the relationship between preinfection hospital exposure and CDI risk overestimates associated costs.[7,24,36] Unlike visit diagnosis codes, toxin assay results provide a presumptive time to infection that we incorporated into 2 different statistical methods addressing time-dependent bias. When using a case definition of either toxin assay being positive, the measured difference in LOS in the multistate model corresponded closely with the difference seen in the "early HA" stratum of a time-stratified propensity-matched analysis (3.3 vs 3.1 days). This finding suggests that measured differences in this study robustly reflect associated costs of HA-CDI in our patient population. Because estimates for each time-to-infection stratum in the matching analysis differed greatly (Figure 3), time to infection clearly contributed bias to the unstratified analysis (Figure 2), demonstrating how the many studies that ignore this bias[3,5,6] produce inflated estimates. In our dataset, ignoring time-dependent bias would lead to a >2-fold overestimation of CDI-associated LOS. Given our findings, we cautiously interpret the results of meta-analyses that conflate ICD-9 code and toxin assay case definitions and often ignore time-dependent bias.[4–6]

To our knowledge, this is the first study to use machine learning on uncurated EMR data to estimate the local cost of CDI. Our models of CDI risk performed on par with prior models fitted to lower-dimensional data.[23,37,38] Because our models are based on tens of thousands of structured fields in the EMR that require neither chart review nor manual curation beyond masking known CDI-related effects, reanalysis of future data is inexpensive. Starting from exported visit data, the entire analysis runs in several hours on standard desktop computers. Therefore, the effects of new interventions against CDI can be efficiently monitored over time, for example, continually testing whether new treatments actually lower the CDI-associated LOS or quantifying cost savings of new preventive strategies that decrease CDI incidence. Changes in LOS can be extrapolated to approximate economic costs by multiplying by the average cost of extra inpatient days, as LOS is the main contributor to the cost of CDI.[3,21,22,36] In our dataset, using the time-dependency adjusted differences in LOS of 3.1–3.3 days and the national average cost of additional inpatient days for CDI cases,[3] the median cost associated with each case would be approximately $10,600–11,300. This cost is substantial in comparison to the national average price for an inpatient visit, which was approximately $13,000 in 2011.[11] Using the average yearly case load observed in the dataset for toxin assay positive cases, our figures represent an annual accounting cost to Mount Sinai of approximately $1.5 million, not including the opportunity cost of bed occupancy by CDI patients or the impact on infection control resources.[36] In principle, our analysis is generalizable to any HAI where laboratory results recorded in the EMR robustly reflect the incidence of infections.

Our study has several limitations. The analysis was designed conservatively, preferring that models underestimate rather than overestimate CDI-associated changes. For example, we censored all patient visits ending in death; therefore, our results are conditioned on patient survival, although a sensitivity analysis that included 12%–16% additional cases ending in patient death yielded similar quantitative and qualitative results. Additionally, restricting our analysis to 1 index visit per

patient certainly excluded many repeat visits for recurrent CDI, which are known to incur higher costs.[12,13,39] We preferred a relatively simple, fast machine learning technique, elastic net regularized generalized linear models, whereas more advanced techniques might marginally improve propensity model accuracy.

Propensity score matching itself has been criticized for potentially introducing bias via collider variables.[40] However, substantial empirical comparisons of estimates from observational and randomized controlled trial data show that propensity matching often reduces bias.[41] Recent investigations of penalized regression propensity matching also show a reduction in bias.[42,43] We believe our implementation reduced bias because our estimate of the effect of CDI on LOS demonstrated significant deviations from unmatched analyses and concordance with the multistate matching analysis (which did not leverage propensity scores or matching). We also note that propensity-matched estimates offer a conservative effect size, which was the intention of this study.

EMR data have known drawbacks compared to clinical research data, such as limitations in time precision, the sparsity of the data, and increased opportunity for coding error. We did not have structured billing data, so we were unable to characterize the exact relationship between LOS and costs beyond the proportional estimate above. Finally, data for only 1 hospital were available for this study. We provide complete code for our analysis so that it may be reimplemented elsewhere and improved by the community.

In conclusion, 2 independent statistical analyses adjusting for time-dependent bias produced similar results for the CDI-associated change in LOS at Mount Sinai (3.1 and 3.3 days), suggesting that automated methods based on machine learning and uncurated EMR data robustly and conservatively estimate the local cost of an HAI in both LOS and financial terms. This procedure is transparent, reproducible, and inexpensive, suggesting that hospitalists and infection control officers can leverage EMR data to estimate their specific, local costs of HAIs on an ongoing basis rather than relying on widely varying benchmarks published by other institutions.

Address correspondence to Andrew Kasarskis, Icahn Institute and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1498, New York, NY 10029 (andrew.kasarskis@mssm.edu).

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/ice.2017.214.

REFERENCES

1. Leffler DA, Lamont JT. Clostridium difficile. *N Engl J Med* 2015;372:1539–1548.
2. Davies KA, Longshaw CM, Davis GL, et al. Underdiagnosis of *Clostridium difficile* across Europe: The European, multicentre, prospective, biannual, point-prevalence study of *Clostridium difficile* infection in hospitalised patients with diarrhoea (EUCLID). *Lancet Infect Dis* 2014;14:1208–1219.
3. Zimlichman E, Henderson D, Tamir O, et al. Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system. *JAMA Intern Med* 2013;173:2039–2046.
4. Ghantoji SS, Sail K, Lairson DR, DuPont HL, Garey KW. Economic healthcare costs of *Clostridium difficile* infection: a systematic review. *J Hosp Infect* 2010;74:309–318.
5. Zhang S, Palazuelos-Munoz S, Balsells EM, Nair H, Chit A, Kyaw MH. Cost of hospital management of *Clostridium difficile* infection in United States—a meta-analysis and modelling study. *BMC Infect Dis* 2016;16:447.
6. Gabriel L, Beriot-Mathiot A. Hospitalization stay and costs attributable to *Clostridium difficile* infection: a critical review. *J Hosp Infect* 2014;88:12–21.
7. Stevens VW, Khader K, Nelson RE, et al. Excess length of stay attributable to *Clostridium difficile* infection (CDI) in the acute care setting: a multistate model. *Infect Control Hosp Epidemiol* 2015;36:1–7.
8. Lofgren ET, Cole SR, Weber DJ, Anderson DJ, Moehring RW. Hospital-acquired *Clostridium difficile* infections: estimating all-cause mortality and length of stay. *Epidemiology* 2014;25:570–575.
9. Katz MH. Pay for preventing (not causing) health care-associated infections. *JAMA Intern Med* 2013;173:2046.
10. Dubberke ER, Carling P, Carrico R, et al. Strategies to prevent *Clostridium difficile* infections in acute care hospitals: 2014 update. *Infect Control Hosp Epidemiol* 2014;35:628–645.
11. Cooper Z, Craig S, Gaynor M, Van Reenen J. The price ain't right? Hospital prices and health spending on the privately insured. *NBER* Working Paper No. 21815; 2015.
12. Dubberke ER, Schaefer E, Reske KA, Zilberberg M, Hollenbeak CS, Olsen MA. Attributable inpatient costs of recurrent *Clostridium difficile* infections. *Infect Control Hosp Epidemiol* 2014;35: 1400–1407.
13. Dubberke ER, Reske KA, Olsen MA, McDonald LC, Fraser VJ. Short- and long-term attributable costs of *Clostridium difficile*-associated disease in nonsurgical inpatients. *Clin Infect Dis* 2008;46:497–504.
14. Greco G, Shi W, Michler RE, et al. Costs associated with health care-associated infections in cardiac surgery. *J Am Coll Cardiol* 2015;65:15–23.
15. Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008–2015. Health Information Technology Dashboard website. https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php. Published 2016. Accessed September 21, 2017.

16. Gray BH, Bowden T, Johansen I, Koch S. Electronic health records: an international perspective on "meaningful use." *Issue Brief (Commonw Fund)* 2011;28:1–18.

17. Etheredge LM. A rapid-learning health system. *Health Aff (Millwood)* 2007;26:w107–w118.

18. Dahabreh IJ, Kent DM. Can the learning health care system be educated with observational data? *JAMA* 2014;312:129–130.

19. Pak TR, Kasarskis A. How next-generation sequencing and multiscale data analysis will transform infectious disease management. *Clin Infect Dis* 2015;61:1695–1702.

20. Krumholz HM, Terry SF, Waldstreicher J. Data acquisition, curation, and use for a continuously learning health system. *JAMA* 2016;316:1669.

21. Wilcox MH, Cunniffe JG, Trundle C, Redpath C. Financial burden of hospital-acquired *Clostridium difficile* infection. *J Hosp Infect* 1996;34:23–30.

22. McGlone SM, Bailey RR, Zimmer SM, et al. The economic burden of *Clostridium difficile*. *Clin Microbiol Infect* 2012;18:282–289.

23. Wiens J, Campbell W. Learning data-driven patient risk stratification models for *Clostridium difficile*. *Open Forum Infect Dis* 2014;1:1–9.

24. Mitchell BG, Gardner A, Barnett AG, Hiller JE, Graves N. The prolongation of length of stay because of *Clostridium difficile* infection. *Am J Infect Control* 2014;42:164–167.

25. Polage CR, Gyorke CE, Kennedy MA, et al. Overdiagnosis of *Clostridium difficile* infection in the molecular test era. *JAMA Intern Med* 2015;175:1–10.

26. Bagdasarian N, Rao K, Malani PN. Diagnosis and treatment of *Clostridium difficile* in adults. *JAMA* 2015;313:398.

27. Moehring RW, Lofgren ET, Anderson DJ. Impact of change to molecular testing for *Clostridium difficile* infection on healthcare facility-associated incidence rates. *Infect Control Hosp Epidemiol* 2013;34:1055–1061.

28. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat* 2011;10:150–161.

29. Li YP, Propert KJ, Rosenbaum PR. Balanced risk set matching. *J Am Stat Assoc* 2001;96:870–882.

30. van Kleef E, Green N, Goldenberg SD, et al. Excess length of stay and mortality due to *Clostridium difficile* infection: a multi-state modelling approach. *J Hosp Infect* 2014;88:213–217.

31. Longtin Y, Paquet-Bolduc B, Gilca R, et al. Effect of detecting and isolating *Clostridium difficile* carriers at hospital admission on the incidence of *C difficile* infections: a quasi-experimental controlled study. *JAMA Intern Med* 2016;176:796–804.

32. Dubberke ER, Reske KA, McDonald LC, Fraser VJ. ICD-9 codes and surveillance for *Clostridium difficile*-associated disease. *Emerg Infect Dis* 2006;12:1576–1579.

33. Scheurer DB, Hicks LS, Cook EF, Schnipper JL. Accuracy of ICD-9 coding for *Clostridium difficile* infections: a retrospective cohort. *Epidemiol Infect* 2007;135:1010–1013.

34. Rhee C, Murphy M V, Li L, Platt R, Klompas M. Improving documentation and coding for acute organ dysfunction biases estimates of changing sepsis severity and burden: a retrospective study. *Crit Care* 2015;19:1–11.

35. Romano PS, Mark DH. Bias in the coding of hospital discharge data and its implications for quality assessment. *Med Care* 1994;32:81–90.

36. Graves N, Harbarth S, Beyersmann J, Barnett A, Halton K, Cooper B. Estimating the cost of health care-associated infections: mind your p's and q's. *Clin Infect Dis* 2010;50: 1017–1021.

37. 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.

38. Tanner J, Khan D, Anthony D, Paton J. Waterlow score to predict patients at risk of developing *Clostridium difficile*-associated disease. *J Hosp Infect* 2009;71:239–244.

39. Rodrigues R, Barber GE, Ananthakrishnan AN. A comprehensive study of costs associated with recurrent *Clostridium difficile* infection. *Infect Control Hosp Epidemiol* 2016:1–7.

40. Pearl J. *Myth, confusion, and science in causal analysis.* Technical Report R-348, University of California website. http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf. Published 2009. Accessed September 21, 2017.

41. Lonjon G, Boutron I, Trinquart L, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Ann Surg* 2014;259:18–25.

42. Athey S, Imbens GW, Wager S. Approximate residual balancing: de-biased inference of average treatment effects in high dimensions. *arXiv.* Cornell University Library website. http://arxiv.org/abs/1604.07125. Published 2016. Accessed September 21, 2017.

43. Antonelli J, Cefalu M, Palmer N, Agniel D. Doubly robust matching estimators for high dimensional confounding adjustment. *arXiv.* Cornell University Library website. http://arxiv.org/abs/1612.00424. Published 2016. Accessed September 21, 2017.

44. Halpin T, Morgan T. *Information modeling and relational databases.* 2nd ed. Elsevier Science; 2010.