# Title

Estimating local costs associated with *Clostridium difficile* infection: a retrospective cohort study using machine learning and electronic medical records

## Authors

Theodore R. Pak, AB,[[1]](#footnote-2) Kieran Chacko, BA, BSc,a Timothy O’Donnell, BS,a Shirish Huprikar, MD,[[2]](#footnote-3) Harm van Bakel, PhD,a Andrew Kasarskis, PhD,a# and Erick R. Scott, MD, MHSa

**Manuscript word count:** 3,050

# Corresponding author: 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, USA; telephone, +1 (212) 659-8542; email [andrew.kasarskis@mssm.edu](mailto:andrew.kasarskis@mssm.edu)

# Research in context

## Evidence before this study

We searched PubMed with the terms difficile AND (cost OR "length of stay") for articles published before December 1, 2016. We prioritized articles that were meta-analyses of the costs of *Clostridium difficile* infection (CDI). We then reviewed the references of papers identified by this strategy and subsequent citations (identified with Google Scholar). Among previous studies, there was wide variation in the definition of CDI (either by ICD-9 codes, toxin assay results, or other criteria), and roughly half of the studies did not use any sort of matching strategy to account for different pre-existing CDI risk in control and cohort populations. Nearly all studies failed to account for bias in the estimated changes in length of stay introduced by differences in time-to-infection. The studies showed substantial variation in the per-patient cost of CDI ($2,871 to $122,318), and furthermore, studies that applied identical methods to multiple hospitals found up to ten-fold variation in per-patient costs of CDI. This suggested that hospitals allocating resources to infection control would benefit from estimating CDI costs in their own patient population, rather than relying on the conflicting estimates in the literature.

## Added value of this study

Traditionally, estimating costs of healthcare-associated infections has required manual review of patient and billing data followed by regression or matching to adjust for covariates. By contrast, we present an automated method for transparently and efficiently estimating changes in length of stay associated with CDI using only the structured fields in electronic medical records. We demonstrate the method on data for 171,938 visits over a 7-year period to an urban tertiary-care hospital. In our dataset, the method used 23,968 variables recorded within the first 24 hours of each visit to automatically match case and control patients by their propensity for CDI as predicted by a machine-learning algorithm. Performance of these CDI risk prediction models was on par with literature reports for prior models run on lower-dimensional, hand curated datasets. We found that our method, after stratifying by time-to-infection, produces a similar estimate for the change in length of stay as a nonparametric multistate model of time-dependent competing risks between CDI and discharge across all time-to-infection intervals. Our methods run within several hours on a standard desktop computer. All implementing software, which requires only open source components, has been made freely available online for other hospitals to use on their own data.

## Implications of all the available evidence

Automated methods based on machine learning and uncurated electronic medical record data are able to robustly estimate the local cost of CDI. Given the prevalence of electronic medical record systems, hospitals can avoid the expense of manually curating datasets to estimate their costs for healthcare associated infections and instead efficiently calculate estimates from the data that is already available. Because these methods are cheap to run on additional data in the future, infection control officers can quickly assess the efficacy of new interventions. More broadly, our study highlights the benefit of fully leveraging electronic medical records to provide localized, robust evidence for decision-making in infection control and clinical infectious diseases.

# Abstract

**Background:** Reported per-patient costs of *Clostridium difficile* infection (CDI) vary by two orders of magnitude among different hospitals, suggesting that precise, local analyses are needed to guide decision-making. We sought to estimate changes in length of stay (LOS) associated with CDI at one hospital using only automatically extractable electronic medical record (EMR) data.

**Methods:** Retrospective cohort study of 171,938 visit records spanning a 7-year period. 23,968 variables were extracted from EMR data recorded within 24 hours of each admission to train an elastic net regularized logistic regression model for propensity score matching. To address time-dependent bias (reverse causation), we stratified comparisons by time-of-infection and fit multistate models.

**Findings:** 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).

**Interpretation:** Changes in LOS can be extrapolated to a marginal dollar cost per CDI case by multiplying by the average cost of an inpatient-day. Infection control officers can leverage automatically extractable EMR data to estimate costs of CDI at their specific institution.

**Funding:** Icahn Institute for Genomics and Multiscale Biology; NIH/NIAID.

# Introduction

*Clostridium* *difficile* infection (CDI) is the most frequently reported healthcare-associated infection (HAI) in the US\autocite{Leffler2015} and the major infective cause of nosocomial diarrhea in developed countries,\autocite{Davies2014} incurring billions of dollars in excess medical costs per year.\autocite{Zimlichman2013} 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.\autocite{Gabriel2014,Ghantoji2010,Zhang2016} Given the high hospital-to-hospital variability of these costs,\autocite{Lofgren2014,Stevens2015} infection control officers, hospital administrators, and clinicians would benefit from estimates tailored to their particular population and healthcare practices. Concretely defining the potential economic savings of CDI prevention would empower stakeholders to prudently choose among the many available validated interventions.\autocite{Dubberke2014,Katz2013}

Measuring costs within healthcare systems is notoriously difficult (particularly in the US), as many hospitals do not have access to structured, itemized reimbursement data linked to all of their patient medical records.\autocite{Cooper2015} Even the institutions that have informatics capabilities to retrospectively link these data have relied on the curation of select variables and chart review to estimate attributable CDI cost.\autocite{Dubberke2008, 2014,Greco2015} Nevertheless, electronic medical record (EMR) systems are used by the majority of first-world acute care facilities.\autocite{Gray2011,Henry2016} 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.\autocite{Dahabreh2014,Etheredge2007,Pak2015} 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.\autocite{Krumholz2016} 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.\autocite{McGlone2012,Wilcox1996,Zimlichman2013}

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 use all available structured data recorded within 24 hours of admission in the EMR—including over 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, avoiding 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.\autocite{Wiens2014} We then use these trained CDI risk models for propensity score matching, which allows estimation of changes in LOS associated with CDI. Most previous studies of CDI cost do not account for the possibility that longer LOS increases the risk of CDI, i.e., reverse causation, and therefore likely overestimate the cost of CDI.\autocite{Mitchell2014,Stevens2015} To adjust for this, we stratify 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 compare 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, NY. Records of adult inpatient visits were extracted from warehoused Epic EMR data and de-identified using the HIPAA Safe Harbor method, 45 CFR §164.514(b)(2). Data was collected on demographics, LOS, time of death, admission sources, reported medications, and the presence of a “008.45” ICD-9 principal or secondary visit diagnosis code denoting “Intestinal infection due to *Clostridium difficile.*” Furthermore, all records of medications administered, abnormal lab result codes, surgery procedure codes, or problem list ICD-9 codes within the first 24 hours after admission were collected as boolean variables (presence or absence). All codes and variables that were uniform across the study population were dropped from the dataset. The relationship between collected data elements, including the maximal cardinality of each datatype, are summarized in Figure 1A. This study was approved by Mount Sinai’s Institutional Review Board as exempt research.

## 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. 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 vs. positive laboratory tests to define CDI cases\autocite{Gabriel2014,Zhang2016} and identify differing positive predictive values for immunoassay and nucleic acid based laboratory tests.\autocite{Bagdasarian2015,Moehring2013,Polage2015} To ensure maximally robust results and allow comparison with prior studies, we repeated our analysis for five definitions of CDI:

1. An “008.45” principal or secondary ICD-9 visit diagnosis code
2. ≥1 positive stool toxin enzyme immunoassay (EIA) lab result
3. ≥1 positive stool toxin polymerase chain reaction (PCR) lab result
4. Either ii or iii
5. Any of i, ii, or iii

Our study’s time range included both a period where the EIA assay was the standard hospital laboratory test (~3 years) followed by a period where the PCR assay was standard (~4 years). For case cohorts (ii) and (iii), comparisons were only permitted with controls from the time range during which that same test was standard. The hospital laboratory only performs toxin assays on unformed stool samples, implying the presence of diarrhea for positive results.

## 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 five 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.\autocite{Austin2011} Matching was repeated using the matched controls against remaining unmatched controls to create a matched-again cohort, testing whether matching alone associates 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 tested with the two-sided Mann-Whitney *U*. Nonparametric Kaplan-Meier estimates of the time-dependent risk of discharge were plotted for matched cohorts.

To address the possible effect of time-to-infection on CDI risk and measured LOS differences, we repeated the analysis for case definition (iv) stratified by the time of the first positive toxin assay, using three 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.\autocite{Li2001} LOS comparisons followed the same procedure as above. We furthermore fit a nonparametric multistate model consistent with previous studies,\autocite{Mitchell2014,Stevens2015,van Kleef2014} under which the mean excess LOS was estimated as the average difference in LOS between patients that had or 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, and all software code for the analysis is available at: <https://github.com/powerpak/cdi-cost>

## Role of the funding source

The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study, and had final responsibility for the decision to submit the publication.

# Results

371,622 records of visits during the study time range were queried from the EMR, with 23,968 variables extracted for each visit (Figure 1). After filtering for the index visit per adult patient and excluding deaths and invalid dates, 171,938 visits were eligible for inclusion and classified into five overlapping case definitions for CDI. Case cohort sizes before matching and overlaps are depicted in Figure S2.

Regularized logistic regression models predicting the risk of CDI acquisition were fit to EMR data from the first 24 hours of each admission for each case definition. The cross-validated mean area under the receiving operator characteristic (AUROC), measuring performance of each model, varied only slightly between case definitions (Figure S3): (i) by ICD-9 code, 0·81 (95% confidence interval [CI], 0·79–0·83); (ii) by positive toxin EIA, 0·76 (95% CI, 0·74–0·78), (iii) by positive toxin PCR, 0.84 (95% CI, 0·82–0·85), (iv) by either toxin assay, 0·81 (95% CI, 0·80–0·82); (v) by any of these, 0·82 (95% CI, 0·81–0·83). The number of selected variables in each model ranged from 373 to 1,027 (Table S2).

For each case definition, over 75% of cases were successfully matched by propensity score to controls (Figure S2 and Table 1). The groups are well matched on demographics and propensity scores (all *P* values >0.1 and standardized differences between -0.1 and 0.1; propensity score distributions in Figure S3). 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: (i) by ICD-9 code, 3·1 days (95% CI, 2·2–3·9); (ii) by positive toxin EIA, 10.1 days (95% CI, 7·3–12·2), (iii) by positive toxin PCR, 6·6 days (95% CI, 5·0–8·1), (iv) by either toxin assay, 7·2 days (95% CI, 5·8–8·3); and (v) 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 (matched-again 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-admit 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—if longer pre-infection LOS increases CDI risk, i.e., reverse causation, this leads to overestimates in attributable cost.\autocite{Mitchell2014,Stevens2015} We therefore performed two follow-up analyses to account for this. First, we stratified the LOS comparison by the time of CDI diagnosis for case definition (iv) into 0–3 day, 3–8 day, and ≥8 day case cohorts, training new propensity models for another matched comparison, with similar matching performance (Figure S4). Since 3 days (or 72 hours) is a typical cutoff for differentiating community acquired (CA) from healthcare-associated (HA) CDI,\autocite{Longtin2016,Polage2015} these strata were named “CA,” “early HA,” and “late HA,” respectively. As suspected, stratification revealed a positive correlation between time of diagnosis and the CDI-associated difference in LOS (Figure 3A). The differences in medians were: for CA, 2·5 days (95% CI, 1·2–3·4); early HA, 3·1 days (95% CI, 1·8–4·4); and late HA, 14·0 days (95% CI, 9·9–17·1). All comparisons between matched cases and controls were again strongly statistically significant, and comparisons between matched controls and matched-again controls were not significant (Figure 3A). Kaplan-Meier plots likewise confirmed a correlation between time of CDI diagnosis and the difference in time-dependent discharge risk (Figure 3B-D).

To further address reverse causation, we fit a multistate model similar to previously published studies\autocite{Mitchell2014,Stevens2015,van Kleef2014} that explicitly estimates time-dependent, competing risks of transitioning to a CDI-positive state vs. being discharged. Figure 4A depicts the three states and allowed transitions of the model. After fitting the model for the case definitions with a time of diagnosis (ii, iii, and iv), the expected remaining LOS can be compared across cohorts that have already transitioned to the CDI infected state vs. 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: (ii) by positive toxin EIA, 3·0 days (95% CI, 2·0–4·0); (iii) by positive toxin PCR, 3·5 days (95% CI, 2·7–4·5); and (iv) by either toxin assay, 3·3 days (95% CI, 2·6–4·0). Notably, the 95% CI for the difference in cohort (iv) 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 seven 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,\autocite{Gabriel2014,Zhang2016,Zimlichman2013} but changes in the case definition resulted in substantial changes in the estimated differences in LOS. Although two hospitals reported good concordance between ICD-9 codes and CDI toxin assay results,\autocite{Dubberke2006,Scheurer2007} 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 S2). 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 (+) 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.\autocite{Rhee2015,Romano1994}

Treating CDI as a baseline condition by ignoring the relationship between pre-infection hospital exposure and CDI risk overestimates associated costs.\autocite{Graves2010,Mitchell2014,Stevens2015} Unlike visit diagnosis codes, toxin assay results provide a presumptive time-to-infection that we incorporated into two 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 suggests that measured differences in this study robustly reflect associated costs of HA-CDI in our patient population. Since 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\autocite{Gabriel2014,Zhang2016,Zimlichman2013} produce inflated estimates. In our dataset, ignoring time-dependent bias would lead to a more than two-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.\autocite{Gabriel2014,Ghantoji2010,Zhang2016}

To our knowledge, this is the first study that uses 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.\autocite{Dubberke2011,Tanner2009,Wiens2014} Since 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, re-analysis 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, e.g., 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 CDI’s cost.\autocite{Graves2010,McGlone2012,Wilcox1996,Zimlichman2013} 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,\autocite{Zimlichman2013} the median cost associated with each case would be approximately $10,600–11,300; this is substantial in comparison to the national average price for an inpatient visit—approximately $13,000 in 2011.\autocite{Cooper2015} Using the average yearly caseload 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.\autocite{Graves2010} In principle, our analysis is generalizable to any HAI where lab results recorded in the EMR robustly reflect the incidence of infections.

Our study has several limitations. The analysis was designed with conservative assumptions, preferring that the models underestimate rather than overestimate changes associated with CDI. For example, restricting to one index visit per patient certainly excluded many repeat visits for recurrent CDI, which are known to incur higher costs.\autocite{Dubberke2008, 2014,Rodrigues2016} We preferred a relatively simple machine learning technique, elastic net regularized generalized linear models, because of its transparency in variable selection and unparalleled speed on sparse training data. More advanced techniques might marginally improve propensity model accuracy at the expense of computational complexity and model interpretability. EMR data has known drawbacks compared to clinical research data, such as limitations in time precision, the sparsity of the data, and increased opportunity for coding error. Also, retrospective matching-based studies face known trade-offs in estimating actual effects of HAIs,\autocite{Graves2010} so we used two separate statistical approaches to control for time-dependent bias, ultimately finding congruous results. We did not have structured billing data, so we cannot trace itemized costs and characterize the exact relationship between LOS and costs beyond the proportional estimate above. Finally, only one hospital’s data was used to implement this analysis, which would benefit from comparison to other hospitals’ data. Therefore, we provide complete code for our analysis so that it may be re-implemented elsewhere and improved by the community.

# Conclusions

Two 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.

# Notes

## Contributors

TRP, KC, AK, and ERS designed the study and developed the models. TRP, KC, TO, SH, HVB, AK, and ERS acquired the data and participated in data analysis. SH reviewed selected variables for the propensity model. TO provided technical support, and HVB, SH, and AK provided administrative and material support. AK and ERS supervised the study. TRP drafted the manuscript. All authors read and approved the final manuscript.

## Declaration of interests

We declare no competing interests.

## Acknowledgements

This work was supported by the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai, and also in part by the NIH/NIAID grants F30AI122673 (TRP), T32GM007280 (TRP), R01AI119145 (HV, KC), and the resources and expertise of the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. We thank Deena Altman, Camille Hamula, and Gopi Patel for their assistance in improving the design of the study and reviewing the manuscript.

# References

{Austin2011} ‘Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies’; Pharmaceutical Statistics, Vol. 10, No. 2 (2011), pp. 150–161. https://doi.org/10.1002/pst.433

{Bagdasarian2015} Rao, K., Malani, P. N.: ‘Diagnosis and Treatment of Clostridium difficile in Adults’; JAMA, Vol. 313, No. 4 (2015), p. 398. https://doi.org/10.1001/jama.2014.17103

{Cooper2015} Craig, S., Gaynor, M., Van Reenen, J.: ‘The Price Ain’t Right? Hospital Prices and Health Spending on the Privately Insured’; NBER Working Paper (2015), p. 21815. https://doi.org/10.3386/w21815

{Dahabreh2014} Kent, D. M.: ‘Can the learning health care system be educated with observational data?’; JAMA, Vol. 312, No. 2 (2014), pp. 129–30. https://doi.org/10.1001/jama.2014.4364

{Davies2014} Longshaw, C. M., Davis, G. L., Bouza, E., Barbut, F., Barna, Z., Delm??e, M., 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)’; The Lancet Infectious Diseases, Vol. 14, No. 12 (2014), pp. 1208–1219. https://doi.org/10.1016/S1473-3099(14)70991-0

{Dubberke2014} Carling, P., Carrico, R., Donskey, C. J., Loo, V. G., McDonald, L. C., Maragakis, L. L., et al.: ‘Strategies to Prevent Clostridium difficile Infections in Acute Care Hospitals: 2014 Update’; Infection Control and Hospital Epidemiology, Vol. 35, No. 6 (2014), pp. 628–645. https://doi.org/10.1086/676023

{Dubberke2006} Reske, K. A., McDonald, L. C., Fraser, V. J.: ‘ICD-9 codes and surveillance for Clostridium difficile-associated disease’; Emerging Infectious Diseases, Vol. 12, No. 10 (2006), pp. 1576–1579. https://doi.org/10.3201/eid1210.060016

{Dubberke2008} Reske, K. A., Olsen, M. A., McDonald, L. C., Fraser, V. J.: ‘Short- and Long-Term Attributable Costs of Clostridium difficile-Associated Disease in Nonsurgical Inpatients’; Clinical Infectious Diseases, Vol. 46, No. 4 (2008), pp. 497–504. https://doi.org/10.1086/526530

{Dubberke2014} Schaefer, E., Reske, K. A., Zilberberg, M., Hollenbeak, C. S., Olsen, M. A.: ‘Attributable Inpatient Costs of Recurrent *Clostridium difficile* Infections’; Infection Control and Hospital Epidemiology, Vol. 35, No. 11 (2014), pp. 1400–1407. https://doi.org/10.1086/678428

{Dubberke2011} Yan, Y., Reske, K. A., Butler, A. M., Doherty, J., Pham, V., Fraser, V. J.: ‘Development and validation of a Clostridium difficile infection risk prediction model.’; Infection Control and Hospital Epidemiology, Vol. 32, No. 4 (2011), pp. 360–6. https://doi.org/10.1086/658944

{Etheredge2007} ‘A rapid-learning health system.’; Health Affairs (Project Hope), Vol. 26, No. 2 (2007), pp. w107-18. https://doi.org/10.1377/hlthaff.26.2.w107

{Gabriel2014} Beriot-Mathiot, A.: ‘Hospitalization stay and costs attributable to Clostridium difficile infection: A critical review’; Journal of Hospital Infection, Vol. 88, No. 1 (2014), pp. 12–21. https://doi.org/10.1016/j.jhin.2014.04.011

{Ghantoji2010} Sail, K., Lairson, D. R., DuPont, H. L., Garey, K. W.: ‘Economic healthcare costs of Clostridium difficile infection: A systematic review’; Journal of Hospital Infection, Vol. 74, No. 4 (2010), pp. 309–318. https://doi.org/10.1016/j.jhin.2009.10.016

{Graves2010} 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.’; Clinical Infectious Diseases, Vol. 50, No. 7 (2010), pp. 1017–1021. https://doi.org/10.1086/651110

{Gray2011} Bowden, T., Johansen, I., Koch, S.: ‘Electronic health records: an international perspective on “meaningful use”.’; Issue Brief (Commonwealth Fund), Vol. 28, No. November (2011), pp. 1–18.

{Greco2015} Shi, W., Michler, R. E., Meltzer, D. O., Ailawadi, G., Hohmann, S. F., Thourani, V. H., et al.: ‘Costs associated with health care-associated infections in cardiac surgery’; Journal of the American College of Cardiology, Vol. 65, No. 1 (2015), pp. 15–23. https://doi.org/10.1016/j.jacc.2014.09.079

{Halpin2010} Morgan, T.: ‘Information Modeling and Relational Databases’; (2nd ed.). Elsevier Science (2010).

{Henry2016} Pylypchuk, Y., Searcy, T., Patel, V.: ‘Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015’; ONC Data Brief (2016), p. no.35.

{Katz2013} ‘Pay for preventing (not causing) health care-associated infections.’; JAMA Internal Medicine, Vol. 173, No. 22 (2013), p. 2046. https://doi.org/10.1001/jamainternmed.2013.9754

{Krumholz2016} Terry, S. F., Waldstreicher, J.: ‘Data Acquisition, Curation, and Use for a Continuously Learning Health System’; JAMA, Vol. 316, No. 16 (2016), p. 1669. https://doi.org/10.1001/jama.2016.12537

{Leffler2015} Lamont, J. T.: ‘Clostridium difficile.’; The New England Journal of Medicine, Vol. 372 (2015), pp. 1539–48. https://doi.org/10.1056/NEJMra1403772

{Li2001} Propert, K. J., Rosenbaum, P. R.: ‘Balanced Risk Set Matching’; Journal of the American Statistical Association, Vol. 96, No. October 2014 (2001), pp. 870–882. https://doi.org/10.1198/016214501753208573

{Lofgren2014} Cole, S. R., Weber, D. J., Anderson, D. J., Moehring, R. W.: ‘Hospital-Acquired Clostridium difficile Infections’; Epidemiology, Vol. 25, No. 4 (2014), pp. 570–575. https://doi.org/10.1097/EDE.0000000000000119

{Longtin2016} Paquet-Bolduc, B., Gilca, R., Garenc, C., Fortin, E., Longtin, J., Trottier, S., 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 Internal Medicine, Vol. 176, No. 6 (2016), pp. 796–804. https://doi.org/10.1001/jamainternmed.2016.0177

{McGlone2012} Bailey, R. R., Zimmer, S. M., Popovich, M. J., Tian, Y., Ufberg, P., Muder, R. R., Lee, B. Y.: ‘The economic burden of Clostridium difficile’; Clinical Microbiology and Infection, Vol. 18, No. 3 (2012), pp. 282–289. https://doi.org/10.1111/j.1469-0691.2011.03571.x

{Mitchell2014} Gardner, A., Barnett, A. G., Hiller, J. E., Graves, N.: ‘The prolongation of length of stay because of Clostridium difficile infection’; American Journal of Infection Control, Vol. 42, No. 2 (2014), pp. 164–167. https://doi.org/10.1016/j.ajic.2013.07.006

{Moehring2013} Lofgren, E. T., Anderson, D. J.: ‘Impact of Change to Molecular Testing for Clostridium difficile Infection on Healthcare Facility-Associated Incidence Rates.’; Infection Control and Hospital Epidemiology : The Official Journal of the Society of Hospital Epidemiologists of America, Vol. 34, No. 10 (2013), pp. 1055–61. https://doi.org/10.1086/673144

{Pak2015} Kasarskis, A.: ‘How Next-Generation Sequencing and Multiscale Data Analysis Will Transform Infectious Disease Management’; Clinical Infectious Diseases, Vol. 61, No. 11 (2015), pp. 1695–1702. https://doi.org/10.1093/cid/civ670

{Polage2015} Gyorke, C. E., Kennedy, M. A., Leslie, J. L., Chin, D. L., Wang, S., Nguyen, H. H., et al.: ‘Overdiagnosis of Clostridium difficile Infection in the Molecular Test Era.’; JAMA Internal Medicine, Vol. 175, No. 11 (2015), pp. 1–10. https://doi.org/10.1001/jamainternmed.2015.4114

{Rhee2015} 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’; Critical Care, Vol. 19, No. 1 (2015), pp. 1–11. https://doi.org/10.1186/s13054-015-1048-9

{Rodrigues2016} Barber, G. E., Ananthakrishnan, A. N.: ‘A Comprehensive Study of Costs Associated With Recurrent Clostridium difficile Infection’; Infection Control & Hospital Epidemiology (2016), pp. 1–7. https://doi.org/10.1017/ice.2016.246

{Romano1994} Mark, D. H.: ‘Bias in the coding of hospital discharge data and its implications for quality assessment.’; Medical Care, Vol. 32, No. 1 (1994), pp. 81–90.

{Scheurer2007} Hicks, L. S., Cook, E. F., Schnipper, J. L.: ‘Accuracy of ICD-9 coding for Clostridium difficile infections: a retrospective cohort.’; Epidemiology and Infection, Vol. 135, No. 6 (2007), pp. 1010–3. https://doi.org/10.1017/S0950268806007655

{Stevens2015} Khader, K., Nelson, R. E., Jones, M., Rubin, M. a, Brown, K. a, Evans, M. E., et al.: ‘Excess Length of Stay Attributable to Clostridium difficile Infection (CDI) in the Acute Care Setting: A Multistate Model.’; Infection Control and Hospital Epidemiology, Vol. 36, No. Cdi (2015), pp. 1–7. https://doi.org/10.1017/ice.2015.132

{Tanner2009} Khan, D., Anthony, D., Paton, J.: ‘Waterlow score to predict patients at risk of developing Clostridium difficile-associated disease’; Journal of Hospital Infection, Vol. 71, No. 3 (2009), pp. 239–244. https://doi.org/10.1016/j.jhin.2008.11.017

{van Kleef2014} Green, N., Goldenberg, S. D., Robotham, J. V., Cookson, B., Jit, M., Edmunds, W. J., Deeny, S. R.: ‘Excess length of stay and mortality due to Clostridium difficile infection: A multi-state modelling approach’; Journal of Hospital Infection, Vol. 88, No. 4 (2014), pp. 213–217. https://doi.org/10.1016/j.jhin.2014.08.008

{Wiens2014} Campbell, W.: ‘Learning Data-Driven Patient Risk Stratification Models for Clostridium difficile’; Open Forum Infectious Diseases, Vol. 1, No. 2 (2014), pp. 1–9. https://doi.org/10.1093/ofid/ofu045

{Wilcox1996} Cunniffe, J. G., Trundle, C., Redpath, C.: ‘Financial burden of hospital-acquired Clostridium difficile infection.’; The Journal of Hospital Infection, Vol. 34, No. 1 (1996), pp. 23–30. https://doi.org/10.1016/S0195-6701(96)90122-X

{Zhang2016} Palazuelos-Munoz, S., Balsells, E. M., Nair, H., Chit, A., Kyaw, M. H., Crobach, M., et al.: ‘Cost of hospital management of Clostridium difficile infection in United States—a meta-analysis and modelling study’; BMC Infectious Diseases, Vol. 16, No. 1 (2016), p. 447. https://doi.org/10.1186/s12879-016-1786-6

{Zimlichman2013} Henderson, D., Tamir, O., Franz, C., Song, P., Yamin, C. K., Keohane, C., et al.: ‘Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system.’; JAMA Internal Medicine, Vol. 173, No. 22 (2013), pp. 2039–46. https://doi.org/10.1001/jamainternmed.2013.9763

# Tables

**Table 1.** Demographic Characteristics of the Study Population and Matched Cohorts

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | No. (%) | Matched cohorts for each CDI case definition | | | | | |
| i. by ICD-9 code | | | ii. by EIA (+) | | |
| No. (%) | | SMD after matching  (*P* value) | No. (%) | | SMD after matching  (*P* value) |
| Characteristic | **All patients**  (n=171 936) | All controls  (n=171 356) | Matched cases & controlsa (n=489) | All controls (n= 73 647) | Matched cases & controlsa (n=274) |
| Female sex | 101 964 (59) | 101 638 (59) | 278 (57) | 0 (1) | 44 132 (60) | 145 (53) | 0 (1) |
| Ageb |  |  |  |  |  |  |  |
| 18-29 | 22 344 (13) | 22 266 (13) | 69 (14) | 0.016 (0.86) | 9 552 (13) | 22 (8) | 0.018 (0.79) |
| 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) |

*(Table 1 cont.)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Matched cohorts for each CDI case definition (cont.) | | | | | |
| iii. by PCR (+) | | | iv. by EIA or PCR (+) | | |
| No. (%) | | SMD after matching  (*P* value) | No. (%) | | SMD after matching  (*P* value) |
| Characteristic | All controls (n=97 351) | Matched cases & controlsa (n=493) | All controls (n=170 994) | Matched cases & controlsa (n=788) |
| Female sex | 57 340 (59) | 254 (52) | 0 (1) | 101 469 (59) | 408 (52) | 0 (1) |
| Ageb |  |  |  |  |  |  |
| 18-29 | 12 714 (13) | 47 (10) | 0.005 (0.99) | 22 265 (13) | 79 (9) | 0.003 (0.98) |
| 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) |

*(Table 1 cont.)*

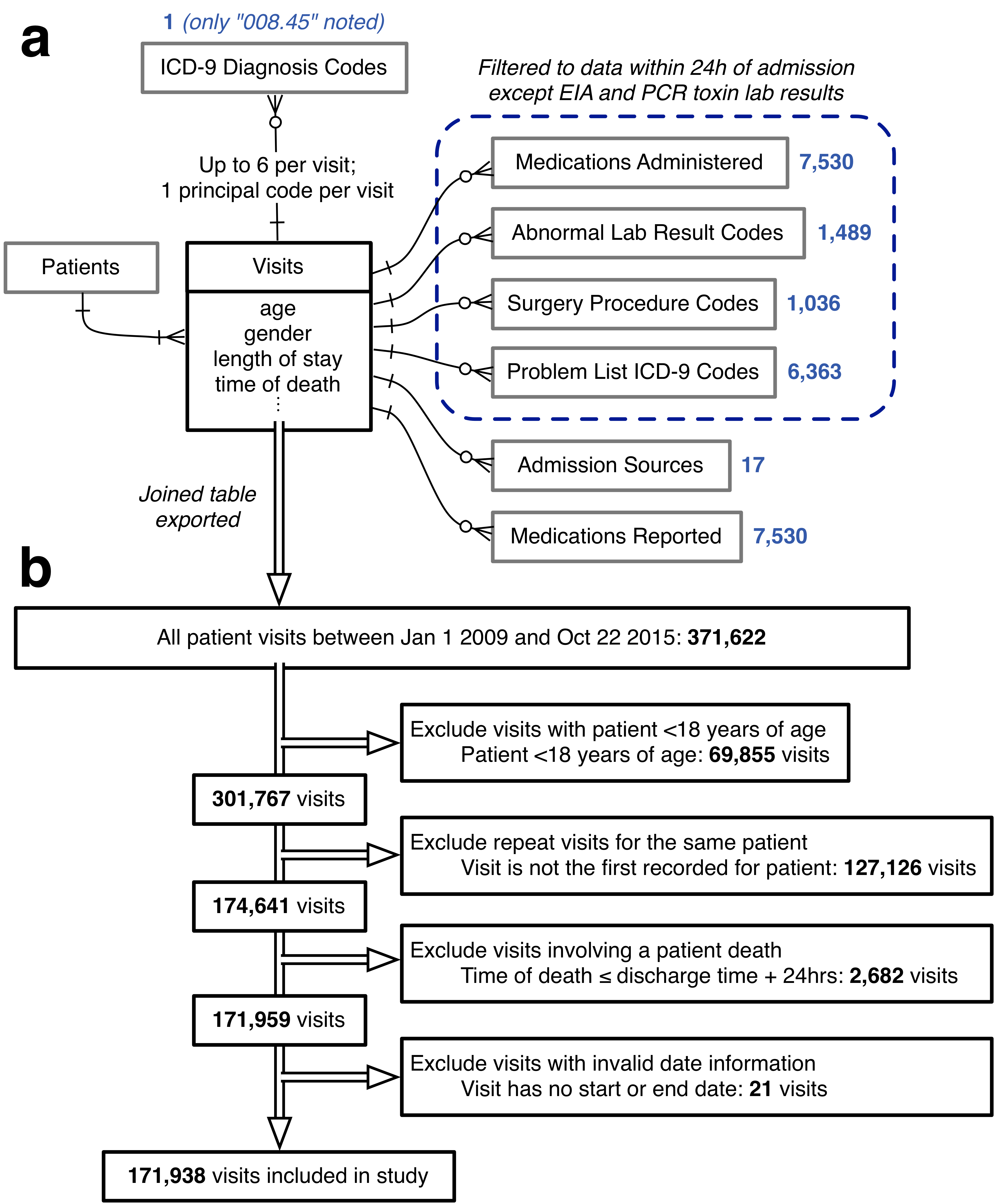
|  |  |  |  |
| --- | --- | --- | --- |
|  | Matched cohorts for each CDI case definition (cont.) | | |
| v. by any of the criteria | | |
| No. (%) | | SMD after matching  (*P* value) |
| Characteristic | All controls (n=170 846) | Matched cases & controlsa (n=945) |
| Female sex | 101 390 (59) | 493 (52) | 0 (1) |
| Ageb |  |  |  |
| 18-29 | 22 245 (13) | 87 (9) | 0.004 (0.93) |
| 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) |

Abbreviation: CDI, *Clostridium difficile* infection; ICD-9, *International Classification of Diseases Ninth Revision*; EIA, enzyme immunoassay; PCR, polymerase chain reaction; SMD, standardized mean difference.

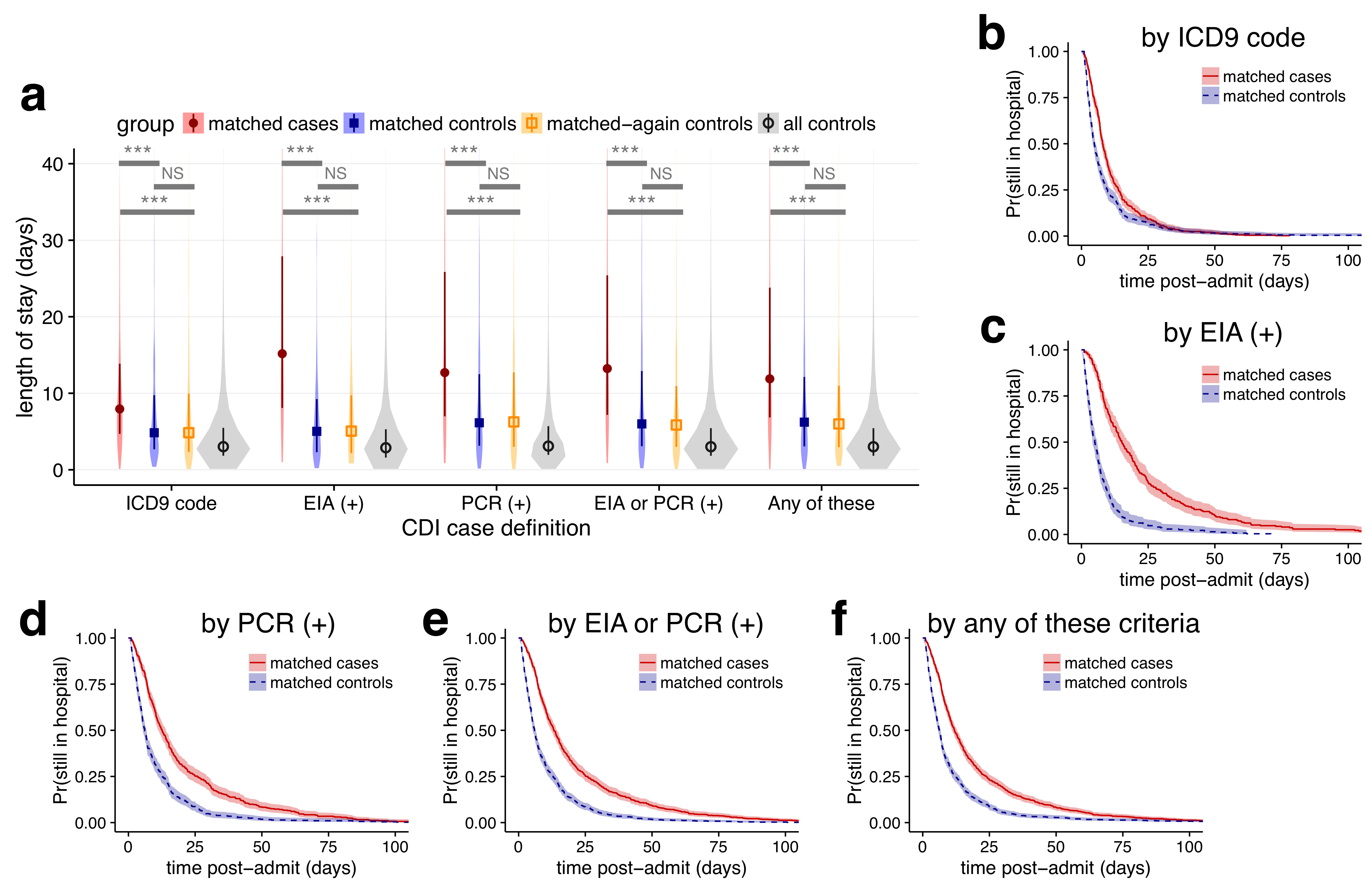
a Separate columns are unnecessary because 1:1 exact matching was performed on the characteristics shown, and therefore all values are identical.

b SMD is shown for age treated as a continuous variable; coarsened exact matching was performed using the listed age ranges.

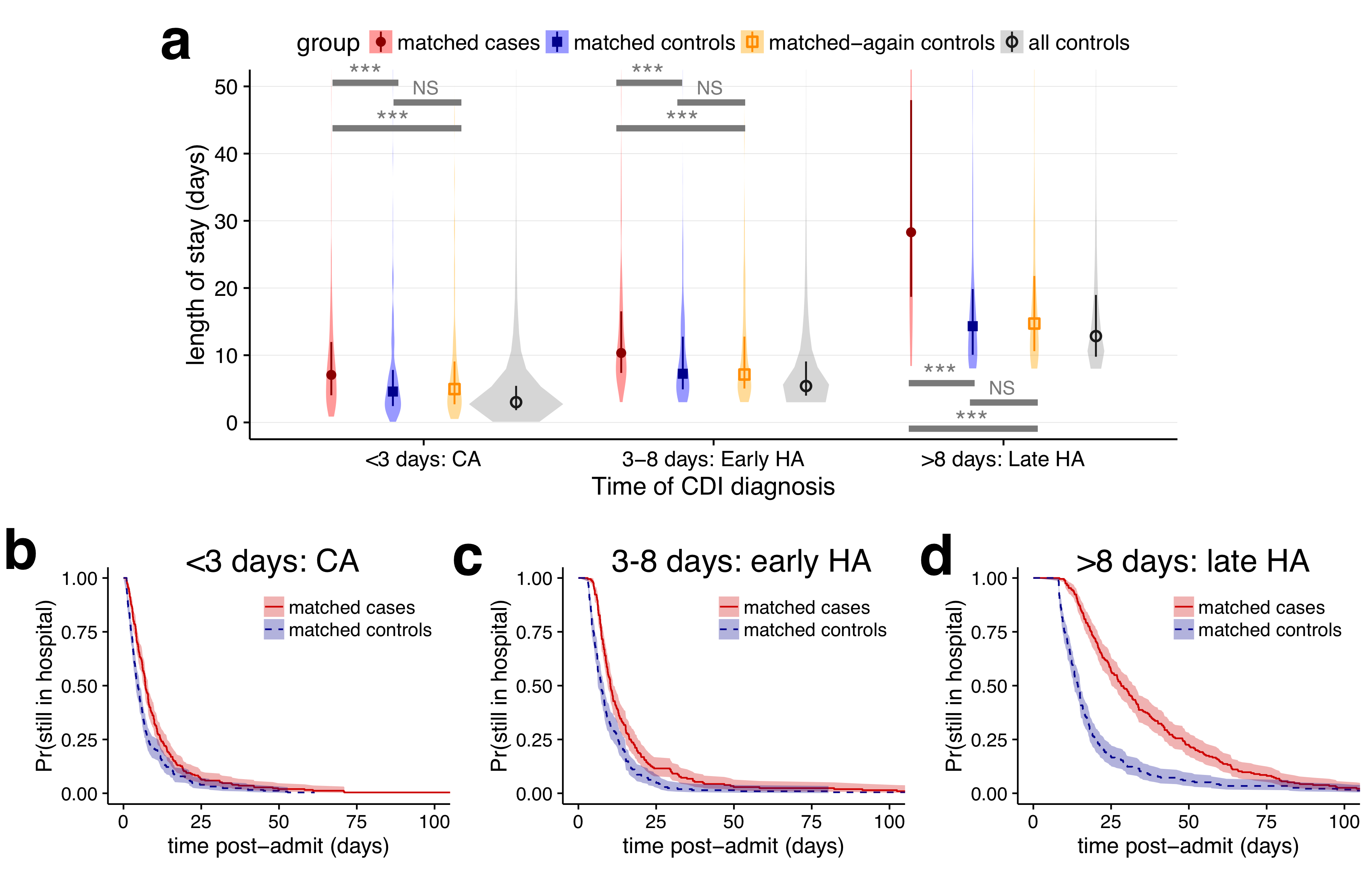
# Figures

****

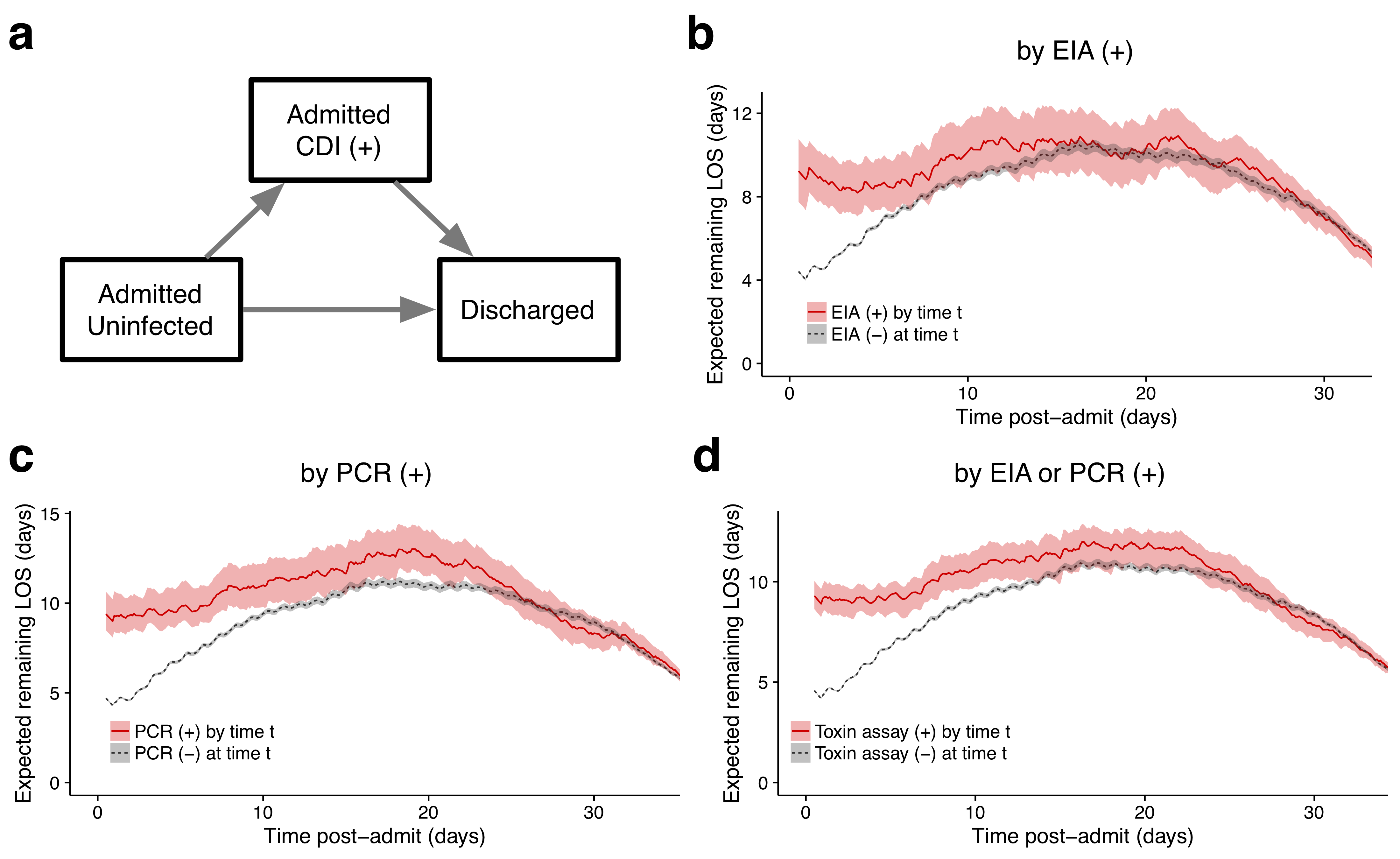
**Figure 1. Data Sources and Inclusion/Exclusion Criteria.** A, entity-relationship diagram for all EMR data used to generate models of CDI propensity, using Information Engineering notation.\autocite{Halpin2010} Boxes represent tables of entities with any directly associated attributes (fields) listed 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 cross-stroke represents “one 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. EMR, electronic medical record; CDI, *Clostridium difficile* infection.

****

**Figure 2. Changes in length of stay for five 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 five 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* < 0.001; NS, not significant [*P* > 0.1]). 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; ICD9, *International Classification of Diseases Ninth Revision*; EIA, enzyme immunoassay; PCR, polymerase chain reaction.

****

**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, matched-again controls, and all controls, for three 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* < 0.001; NS, not significant [*P* > 0.1]). 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 three 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.

****

**Figure 4. Multistate model of expected remaining length of stay for *Clostridium difficile* infection case definitions involving toxin assays.** A, three 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-admit 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.

1. Icahn Institute and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY [↑](#footnote-ref-2)
2. Division of Infectious Diseases, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY [↑](#footnote-ref-3)