BMC Infectious Diseases

RESEARCH

Open Access

Unbiased identification of risk factors for invasive *Escherichia coli* disease using machine learning

Erik Clarke[1], Christel Chehoud[1], Najat Khan[1], Bart Spiessens[2], Jan Poolman[3] and Jeroen Geurtsen[3*]

Abstract

Background Invasive *Escherichia coli* disease (IED), also known as invasive extraintestinal pathogenic *E. coli* disease, is a leading cause of sepsis and bacteremia in older adults that can result in hospitalization and sometimes death and is frequently associated with antimicrobial resistance. Moreover, certain patient characteristics may increase the risk of developing IED. This study aimed to validate a machine learning approach for the unbiased identification of potential risk factors that correlate with an increased risk for IED.

Methods Using electronic health records from 6.5 million people, an XGBoost model was trained to predict IED from 663 distinct patient features, and the most predictive features were identified as potential risk factors. Using Shapley Additive predictive values, the specific relationships between features and the outcome of developing IED were characterized.

Results The model independently predicted that older age, a known risk factor for IED, increased the chance of developing IED. The model also predicted that a history of ≥ 1 urinary tract infection, as well as more frequent and/or more recent urinary tract infections, and ≥ 1 emergency department or inpatient visit increased the risk for IED. Outcomes were used to calculate risk ratios in selected subpopulations, demonstrating the impact of individual or combinations of features on the incidence of IED.

Conclusion This study illustrates the viability and validity of using large electronic health records datasets and machine learning to identify correlating features and potential risk factors for infectious diseases, including IED. The next step is the independent validation of potential risk factors using conventional methods.

Keywords Invasive E. coli disease, Machine learning, Electronic health records, Disease burden per status today

Introduction

Commensal *Escherichia coli*, a gram-negative bacterium, colonizes the gastrointestinal tract soon after birth, establishes a symbiotic relationship, and comprises part of the normal intestinal microbiota in humans [1]. Conversely, pathogenic variants of *E. coli* are divided into intestinal and extraintestinal pathogens and cause various infections, such as urinary tract, inflammatory bowel, diarrheal, and bloodstream infections. Extraintestinal pathogenic *E. coli* (ExPEC) strains can live in the intestinal microbiota without causing disease but can be pathogenic in other body sites [1, 2]. ExPEC strains infect otherwise sterile body sites, such as blood, cerebrospinal fluid, pleural or peritoneal fluid, or renal parenchyma, which may lead to severe and potentially lethal invasive diseases (e.g., bacteremia,

*Correspondence:
Jeroen Geurtsen
jgeurtse@its.jnj.com
[1] Janssen Research and Development Data Sciences, Spring House, PA, USA
[2] Janssen Research and Development, Beerse, Belgium
[3] Janssen Vaccines and Prevention, Leiden, The Netherlands

sepsis, meningitis) [3, 4]. ExPEC strains possess numerous virulence factors that allow strains to survive in different extraintestinal compartments [1, 5–9].

Invasive *E. coli* disease (IED), also known as invasive ExPEC disease, is a leading cause of sepsis and bacteremia in older adults and is frequently resistant to antimicrobial therapy [10–12]. The incidence rates for *E. coli* bacteremia are higher in adults aged ≥ 60 years compared with the general population, increasing up to and beyond 85 years [13–15]. More than 50% of *E. coli* bacteremia cases have a urinary origin [12].

Hospitalizations due to *E. coli*–related urinary tract infections (UTIs), intra-abdominal infections, bacteremia, and sepsis continue to increase, and antimicrobial resistance is one of the contributing factors [7, 11]. The treatment of IED is complicated by multidrug resistance driven by acquisition of plasmid-encoded AmpC β-lactamases, extended-spectrum β-lactamases, and carbapenemases [7]. Multidrug resistance among ExPEC strains contributes to treatment failures with further consequences on hospitalizations, morbidity, and healthcare costs [11].

Certain comorbidities, as well as procedures, increase the risk of developing IED. A study revealed that risk factors for *E. coli* bacteremia in men were urinary catheterization and urinary incontinence; risk factors in women were cancer, chronic renal failure, congestive heart failure, and urinary incontinence [14]. A systematic literature review concluded that renal dialysis, solid-organ transplantation, neoplastic disease, and indwelling vascular and urinary catheters increase the risk for *E. coli* bacteremia, with the authors stressing the need for further research to identify additional risk factors [13].

Traditional approaches to identifying risk factors, such as systematic literature reviews, require significant time and domain expertise. However, the performance of commonly used machine learning approaches, such as linear regression models, support vector machines, and decision trees, has often been surpassed by XGBoost, a regularized gradient-boosting model [16, 17]. In addition, XGBoost can use mixtures of continuous, categorical, and partially missing features that would be challenging to implement in, e.g., regression models. Recent studies have shown that XGBoost models could predict risk factors for COVID-19 critical illness and mortality and multidrug-resistant gram-negative bacilli in patients with hematologic conditions and febrile neutropenia [18–21]. XGBoost-based approaches provide good to high accuracy for predicting risk factors for studied conditions and outcomes and require comparatively fewer resources [16, 18–20].

The objective of this study was to identify and validate potential risk factors for IED that did not rely on existing knowledge of the disease using patient data extracted from the Optum® de-identified Electronic Health Record dataset (Optum® EHR) (released September 8, 2021) [22]. To achieve this, we employed a novel approach that leveraged an XGBoost model trained to predict an individual's likelihood of developing IED (Supplementary Fig. S1), identified key predictive features using Shapley Additive exPlanation (SHAP) values [23], and reinterpreted these SHAP values into simplified exposure variables for further evaluation using traditional epidemiological measures. This method does not require extensive domain knowledge or bespoke feature engineering, is robust to missing data, and produces results that can be validated in independent datasets and checked against the medical literature. In our study, this method corroborated known major risk factors identified in previous work and suggested multiple novel and high-impact potential risk factors for IED.

Methods

Study design

A patient-level prediction model was used to identify potential risk factors for IED in patients within 14 days to 1 year after a patient-specific index date [24] (Supplementary Fig. S1). The index date was defined as the patient's first healthcare encounter between January 1, 2014, and December 31, 2016, and could represent a clinically meaningful occurrence, such as the start of a new treatment, or an arbitrary date, such as a routine office visit or screening. A 14-day washout period was chosen to minimize the chance that acute IED symptoms at the index date would serve as predicting features. Data from ≤ 2 years prior to the index date were used to inform the prediction.

Patients were aged ≥ 18 years at the index date and had 2 years of continuous observations prior to the index date (also defined as the lookback period) and for 1 year following the index date. IED was defined as a diagnosis of *E. coli* sepsis (*International Classification of Diseases, Tenth Revision* A41.51) or a positive *E. coli* culture from a normally sterile body site, such as blood, cerebrospinal fluid, pleural fluid, or peritoneal fluid.

Data source and feature engineering

Data were collected from the Optum de-identified Electronic Health Record dataset (release September 8, 2021) [22] and converted to the Common Data Model schema, version 5.3.1 [25]. Conversion to the Common Data Model allowed the terms from the native database to be mapped to a standardized hierarchical vocabulary of concepts. Optum's data are derived from more than 50 healthcare provider organizations in the United States, which include > 700 hospitals and 7000 clinics, with anonymized longitudinal Optum® EHR data for approximately 99.5 million distinct patients aged 0 to 82 years (individuals > 82 years were censored to prevent de-anonymization). Data include basic demographics, diagnoses of medical conditions, prescriptions, procedures, inpatient and outpatient visits, and laboratory measurements.

Upon identification of the patient cohort, each patient was randomly assigned to either a training (60%), validation (20%), or test (20%) group. Features were then created from each patient's conditions, drugs, procedures, laboratory measurements, age, and biological sex from the 2 years prior to their index date. In addition, higher-level features for classes of conditions, medications, and procedures were added (e.g., "systemic corticosteroids" for prednisone), based on the Medical Dictionary for Regulatory Activities [26], First DataBank Enhanced Therapeutic Classification [27], and Current Procedural Terminology, Fourth Edition [28] vocabularies, respectively, as mapped by the Common Data Model. To select the features that would be used in the model, a cohort with a 1:1 case–control ratio was randomly sampled from the training patients, and the 100 most prevalent features in each category were identified. This resulted in a total of 663 total features, including age and sex.

Age was encoded directly as the number of years of age at index date. Sex was encoded as a binary variable where "True" indicated female. Laboratory measurements were encoded as an ordinal variable based on the most recent laboratory value for a patient relative to their index date, with 1 indicating a value below the laboratory-reported normal range, 2 indicating within normal range, and 3 indicating above normal range. Zero was used to indicate that the patient did not have a measurement recorded during the lookback period. For all other feature groups (conditions, drugs, procedures, and healthcare visits), a weighted sum was calculated based on the number of occurrences of that event in the patient's lookback period. Specifically, the feature value for each patient was calculated as $\sum_{i=1}^{n} e^{-dt_i}$, where n is the total number of observed events for that feature, d represents a decay factor, and t_i represents the number of days between event i and the index date. Patients who had no observed events for a given feature during the lookback period would have a feature value of 0. This encoding allowed information regarding the frequency and recency of a patient's events to be encoded in a single variable. From this pool of features, those with 0 variance or with high correlation with another feature (Pearson $r > 0.9$) were removed.

Model design, tuning, and evaluation

An XGBoost model was trained on the binary prediction task using the training set [16]. XGBoost was used because it is performant on large datasets, can accommodate features of varying scales and missingness, and can learn more flexible relationships between features and outcomes than would be possible using, e.g., regression models. However, XGBoost has multiple configuration parameters (or "hyperparameters") whose optimal values are not known a priori. These values must be identified through a process of hyperparameter tuning; for this study, a grid search across the parameter space was performed on threefold splits of the training dataset. The parameters that resulted in highest mean average precision across the 3 folds were then used in the final model. Average precision (a discrete analogue of the area under the precision-recall curve) was chosen as the evaluation metric because it is robust enough to class imbalances in the data, and the dataset for this study was highly imbalanced (see "Results" section). The final model was tuned on the entirety of the training set, and its performance was evaluated on the test set using average precision, area under the receiver operating characteristic curve, and other metrics.

Model interpretation

Interpretation of the feature-outcome relationships learned by the model is essential for the objective of this study, which is to identify potential risk factors, i.e., features that are highly predictive of the outcome. Beyond a certain threshold, improving the model's performance yields diminishing returns as the major feature-outcome relationships stabilize.

Shapley values are a concept originating from game theory that represent the individual contributions of members of a group to an outcome. SHAP values [23] are a method of applying Shapley values to predictive models and represent the individual contribution of a feature's value to the model's predictions. In other words, the sum of the SHAP values for all values across all features equals the output of the model for a single prediction.

SHAP values were used in this study to both identify important features and to characterize the responses of the model to different values of a given feature. Feature importance was calculated using the mean absolute SHAP value across all patient values for that feature. A high mean absolute SHAP value indicates a feature that was impactful in the model's decisions across the cohort of patients. To identify the most important features to the model, features with the highest mean absolute SHAP value were selected.

However, the mean absolute SHAP value does not characterize the actual relationship between the feature and the outcome learned by the model. To elucidate that relationship for each of the most important features, the feature values and SHAP values across all patients were plotted. These plots were used to illustrate the distribution of feature values for the cohort and their effect on the model's predictions.

Identifying potential risk factors

From manual inspection of these plots, an approximate threshold was identified for each feature that separated impactful values from less impactful values. This threshold was then used to reformulate previously categorical or continuous features into binary exposure variables (e.g., "age in years" might become "age ≥ 30," or "days since hospital visit" might become "hospital visit in past 30 days"). These exposure variables were then used to train a separate logistic regression model of the outcome on the original population. From this model, estimates of the relative risk for each exposure variable were obtained. Exposure variables with high relative risk were identified as potential risk factors for further investigation.

Operational environment

Feature engineering, model tuning, and all other analyses were performed in a computational environment with 30 cores and 230 GiB of RAM, running Ubuntu 18.04.3. Original licensed data were stored in a database server prior to patient selection, after which the data were stored locally.

Results

Study population and prevalence

The study population used to train and test the model comprised 22,041,367 patients from the database who met the previously specified criteria (e.g., age ≥ 18 years at the time of their index date [defined as their first healthcare encounter between January 1, 2014, and December 31, 2016] and ≥ 2 continuous years of observation in the database prior to the index date). Of these, 5362 developed IED within 14 days to 1 year of their index date, for a case prevalence of 0.0002 (or 1 case per 5000 patients).

Model performance

Patients were randomly selected to be in the training (80%) or test (20%) sets, and an XGBoost model was trained on a binary prediction task using the training set. Reflecting the low real-world incidence of IED, the study population was strongly skewed towards non-IED patients, and this skew was reflected in the training data. Therefore, model performance during hyperparameter tuning was measured by its average precision, a discrete analogue of the area under the precision-recall curve, as it is more appropriate for imbalanced data (see "Methods" section). After hyperparameter tuning, the final XGBoost model had an average precision of 0.0031 and a receiver operating characteristic area under the curve of 0.85 on the test dataset (Supplementary Figs. S2a and S2b). Although an average precision of 0.0031 may appear to suggest poor performance, the "no skill" threshold for this metric on this dataset is 0.000243 (the proportion of cases in the total population), which is an order of magnitude lower than the model's value. These metrics indicate that the final model performed substantially better than random at predicting which patients would develop IED, given the imbalance in the data. For this study, the model's performance is primarily useful in understanding whether the features identified by the model reflect actual predictive utility. In other words, there would be no value in the interpretation of important features if the model were unskilled.

Features influential to model predictions

Features derived from the Optum® EHR data for all patients were evaluated based on SHAP values to assess the contribution of each feature to the model's predictions. The features with the greatest mean absolute SHAP value (i.e., those most influential and "important" to the model's predictions) were age, inpatient and emergency department (ED) visits (irrespective of reason for visit), furosemide consumption, type 2 diabetes (T2D) without complication, UTIs, use of laxatives, phlebosclerosis, female sex, outpatient or office visits, other higher-level terms indicating routine care (such as office visits), use of antihistamines, and magnetic resonance imaging clinic visits (Fig. 1a).

Because the training process of XGBoost involves nondeterministic sampling (except for certain combinations of hyperparameters), the calculated SHAP values can vary between models with identical hyperparameters and comparable performance. As a check of the robustness of the identified features, 10 additional models were trained using different random seeds. The mean absolute SHAP value of each feature was calculated for all 10 models and then each was ranked in descending order. Although the mean absolute SHAP value varied between individual models, the ranking for the topmost important features remained stable. For the following analyses, a single model's calculated SHAP values were used.

The use of SHAP values also permitted inspection of the model's encoding of the dependency between a feature and the outcome. In many cases, this relationship was found to be non-linear. For instance, for some features, in patients for whom the feature was never observed, the overall effect on the model's prediction was

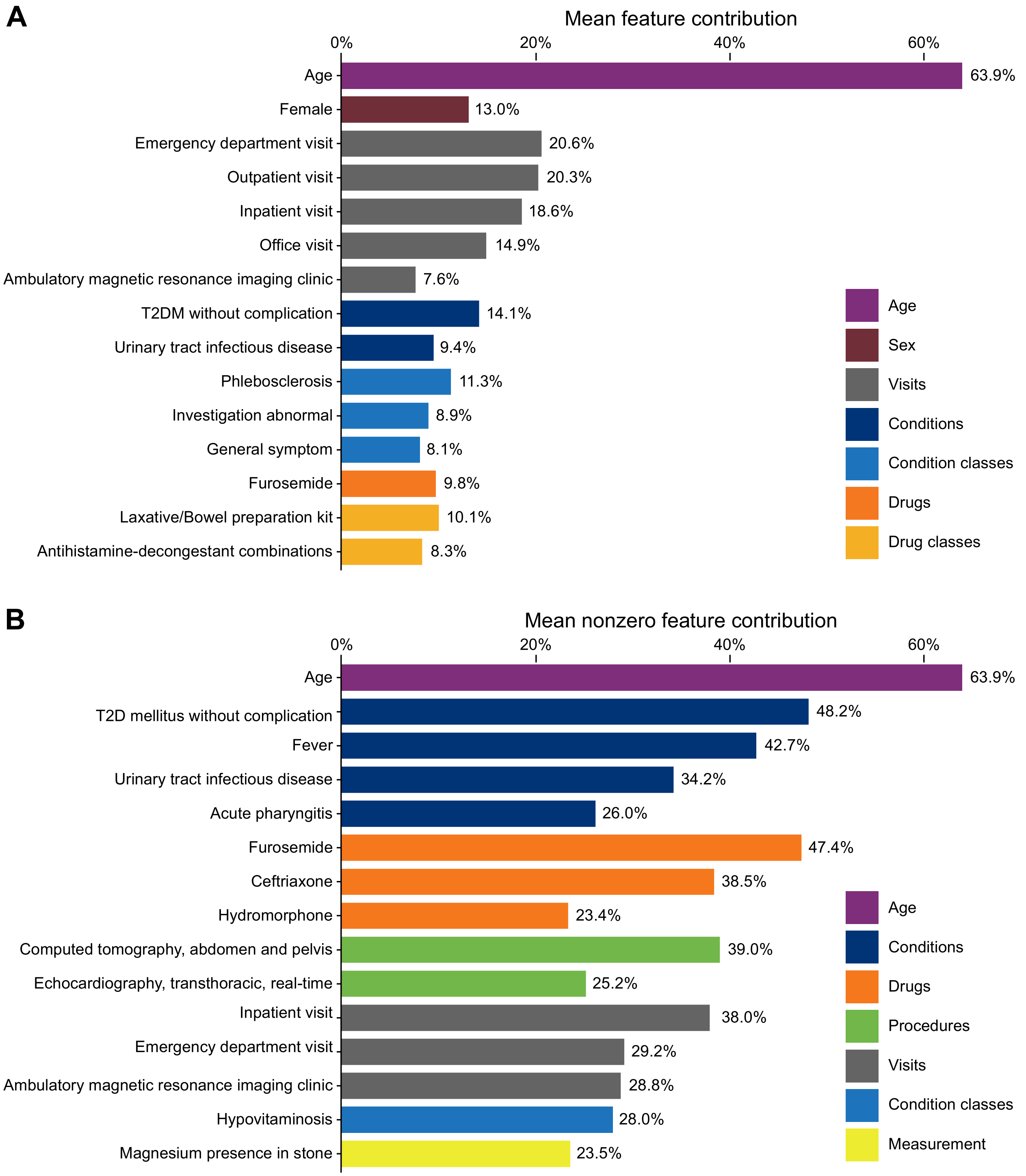

Fig. 1 Most important features to the model predictions; **a** features are ranked by feature type and mean feature contribution (mean absolute SHAP value) across all patients and **b** patients for whom that feature occurred at least once in their lookback period. 2D, 2-dimensional; SHAP, Shapley Additive exPlanationp; T2D, type 2 diabetes

small or negative, but for patients with ≥ 1 occurrence of the feature, the contribution of that feature was substantial and increased with its recency or frequency. As these features may not be classified as important using the mean absolute SHAP value heuristic, especially if the majority of patients do not have that feature, the mean absolute SHAP value among those patients with ≥ 1 occurrence of a feature was also calculated. This revealed a slightly different ranking of features, with ceftriaxone and hydromorphone consumption, fever, echocardiography, computed tomography, and measurements related to kidney stones ranking higher (Fig. 1b).

Age

A patient's age at index was the largest contributor to the model's predictions and showed a strong correlation with relative risk of developing IED (Fig. 2). The relative risk contributed by a patient's age in the model was < 1 for patients aged 18–59 years but > 1 for those aged ≥ 60 years, indicating that the latter age group is at higher risk for IED. Moreover, the relative risk associated with age showed an almost linear increase with age > 60 years.

T2D and UTIs

The presence of certain medical conditions in the 2 years prior to the patient's index date was a substantial contributor to their relative risk for IED, as predicted by the model. A history of T2D or a prior occurrence of a UTI was associated with a relative risk > 1 (Fig. 3). Patients with these conditions in the past 2 years showed an increased risk as the frequency or recency of the events increased. More frequent and/or more recent UTIs were predicted to increase the relative risk between 1.25 and 2.5.

Healthcare visits

Patients with ≥ 1 ED or inpatient visit for any reason in the prior 2 years had an increased risk for developing IED (Fig. 4). Of those with > 1 such visit, increasingly recent or frequent ED or inpatient visits substantially increased their relative risk for IED. Patients who had no recorded outpatient visits in the prior 2 years were also at higher risk for IED. The risk decreased for patients with more frequent or regular outpatient visits. Potential explanations for these associations are outlined in the following sections.

Interpreting model results as exposures

The features described above were also reframed as binary exposure variables (e.g., "aged ≥ 60 years" or "any prior history of UTI") and the relative risk for IED considered between exposed and unexposed groups.

Age

Based on the increase in the model's predicted relative risks for patients aged ≥ 60 years (Fig. 2), an "exposure" was created that divided the study population between those ≥ 60 years at the index date ("exposed" group) and those < 60 years ("unexposed" group). Approximately 30% of the study population fell into the exposed group, with a case prevalence of 0.0005 (n = 6,677,443) compared with 0.0001 among the unexposed (n = 15,357,535). Age ≥ 60 years was associated with a relative risk of 4.90 and an attributable fraction among the exposed (AF_e) of 0.80.

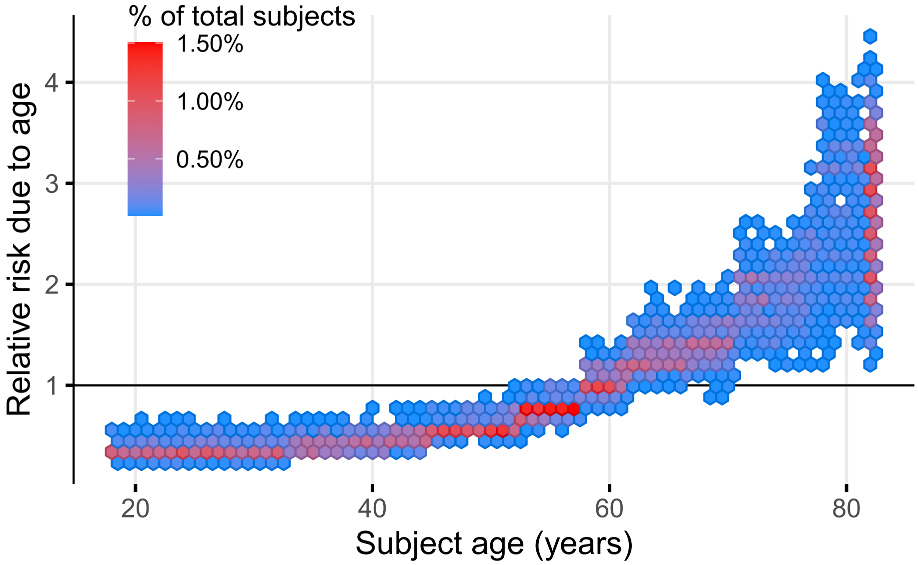

Fig. 2 Effect of age on relative risk of IED. The marginal effect of patient age on that patient's relative risk for IED. The x-axis represents the age of the patient in years, and the y-axis represents the relative risk for IED attributable to the patient's age (as derived from the per-patient SHAP value). Each bin represents patients who share that range of *x*- and *y*-values, with the color indicating how many patients in the total cohort fall into that bin. IED, invasive *Escherichia coli* disease; SHAP, Shapley Additive exPlanation

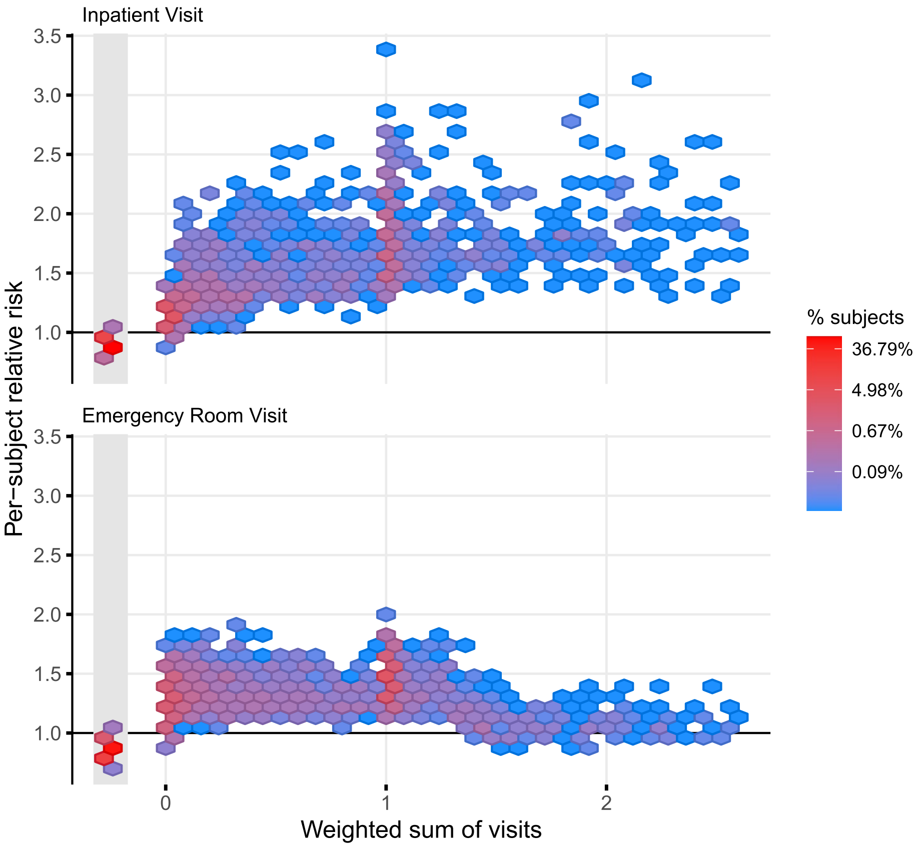

Fig. 3 Effect of T2D and UTIs on risk for IED. The x-axis represents the weighted sum of the number of times a patient had a recorded diagnosis of a UTI or T2D in their lookback period. The y-axis represents the relative risk for IED attributable to the weighted sum of those diagnoses. The color of each bin represents the percentage of patients who fall into that range of x- and y-values. Increasing values of x can be due to either more recent or more frequent diagnoses, or both. Note that especially in the case of chronic diseases, such as T2D, the number of diagnoses likely reflects increased healthcare utilization. IED, invasive *Escherichia coli* disease; T2D, type 2 diabetes

History of UTI

Additionally, the model attributed a higher risk to patients who had ≥ 1 UTI in the past 2 years than those who did not (Fig. 3). Using this as an exposure variable, case prevalence in the exposed group (representing approximately 5% of the study population) was 0.0012 (n = 1,071,192) compared with 0.0002 (n = 20,968,186) in the unexposed group. The relative risk in the exposed group was 6.05, with an AF_e of 0.83.

History of T2D

Approximately 7% of the study population had a record of T2D in the previous 2 years. Using this as an exposure variable, the case prevalence among the exposed group was 0.0009 compared with 0.0002 among the unexposed group, with a relative risk of 5.01 and AF_e of 0.80, lower than the relative risk associated with a history of UTIs.

Inpatient visits

Approximately 10.6% of the study population had ≥ 1 inpatient visit in the past 2 years, with a case prevalence

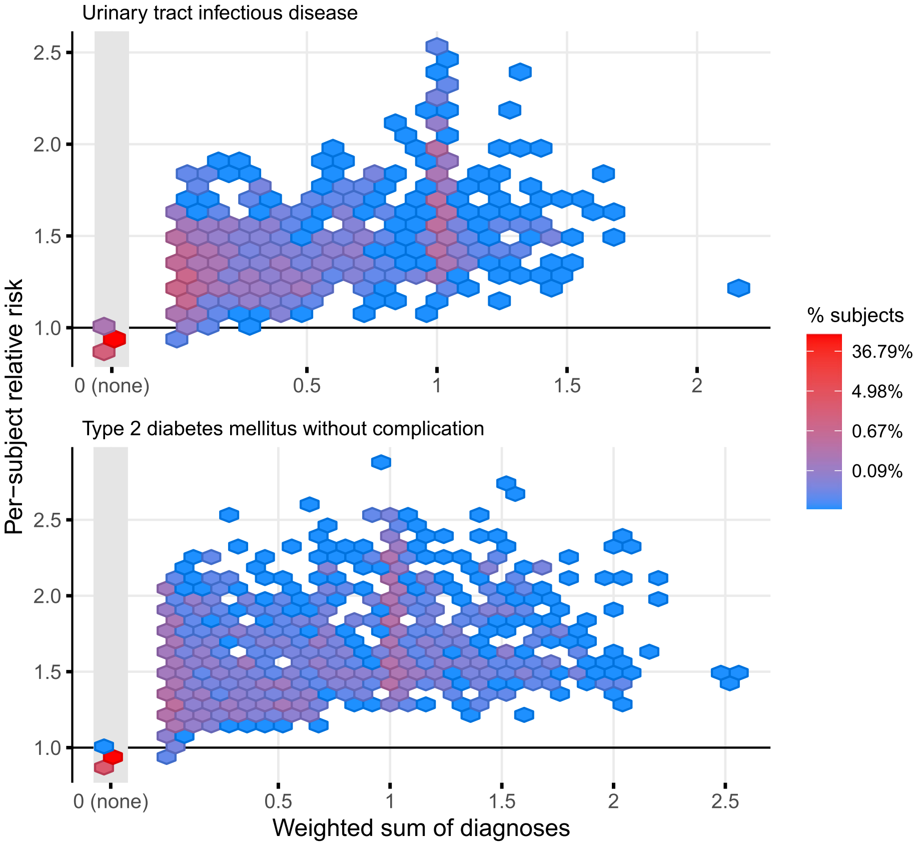

Fig. 4 Effect of healthcare visits on IED prediction; **a** the marginal effect of inpatient visits or **b** ED visits on the relative risk for IED. The x-axis represents the weighted sum of the number of times a patient had a recorded visit of that type in their lookback period, and the y-axis represents the relative risk for IED attributable to the weighted sum of those visits. The color of each bin represents the percentage of patients who fall into that range of *x*- and *y*-values. ED, emergency department; IED, invasive *Escherichia coli* disease

of 0.0010 compared with 0.00016 for those without an inpatient visit. The relative risk for IED in this population was 6.31 (AF_e, 0.84), similar to the relative risk in patients with a history of UTI.

Multivariate exposure variables

The exposures described above were combined to understand whether the high-risk populations they identified were independent from one another (and thus represented separate high-risk groups). Patients aged ≥ 60 years with a history of UTI in the past 2 years represented 1.9% of the population but had nearly 10 times the case prevalence of those who did not meet these criteria (0.0022 vs. 0.0002, respectively). The relative risk in the exposed group was 10.56, with an AF_e of 0.91—substantially higher than either feature considered separately (Table 1).

Adults aged ≥ 60 years with a history of T2D represented 4.3% of the study population and had a relative risk of 6.30 and AF_e of 0.84, comparable to the risk attributable to a history of UTI. This relatively small increase in risk in the combined population compared with either adults ≥ 60 years or those with T2D suggests that they are not independent high-risk groups; indeed, patients with

Table 1 Summary of variables identified to be influential to model predictions based on the mean absolute SHAP value

Univariate exposure variables	**Population,[a]%**	**IED case prevalence[b]**	**Relative risk**	**AF_e**
Age ≥ 60 years	30.3	0.0005	4.90	0.80
History of UTI[c]	4.9	0.0012	6.05	0.83
History of T2D[d]	7.2	0.0009	5.01	0.80
Inpatient visit[e]	10.6	0.0010	6.31	0.84
Multivariate exposure variables	Population, %	IED case prevalence	Relative risk	AF_e
Age ≥ 60 years and history of UTI[c]	1.9	0.0022	10.56	0.91
Age ≥ 60 years and history of T2D[d]	4.3	0.0012	6.30	0.84
Age ≥ 60 years and inpatient visit[e]	4.3	0.0017	9.94	0.90
Age ≥ 60 years and history of UTI and inpatient visit	0.7	0.0044	20.44	0.95
History of UTI and inpatient visit	1.3	0.0030	14.90	0.93

AF_e attributable fraction among the exposed, *IED* invasive *Escherichia coli* disease, *SHAP* Shapley Additive explanation, *T2D* type 2 diabetes, *UTI* urinary tract infection

[a] Patients who were aged ≥ 18 years at index and had ≥ 2 years of observation in the database prior to index

[b] Expressed as the ratio of IED cases within the risk group to the total number of patients in the risk group

[c] Patients who had ≥ 1 UTI in the past 2 years

[d] Patients with a record of T2D in the past 2 years

[e] Percentage of cases that can be assigned to a specific risk factor in patients with ≥ 1 inpatient visit in the past 2 years

a history of T2D in the age group ≥ 60 years had a modestly increased relative risk of 2.89 (AF_e, 0.65) compared with adults ≥ 60 years without T2D. In contrast, patients with a history of UTI among the age group ≥ 60 years had a relative risk of 5.24 (AF_e, 0.81) compared with adults ≥ 60 years without a history of UTI. This is comparable to the relative risk associated with a history of UTI in the general ≥ 18 years study population, suggesting that age and UTIs identify separate high-risk groups.

Patients aged ≥ 60 years who had ≥ 1 inpatient visit showed a relative risk of 9.94 (AF_e, 0.90); those with a prior history of both UTI and inpatient visits had a relative risk of 14.90 (AF_e, 0.93). Patients meeting all 3 criteria (age ≥ 60 years, history of ≥ 1 UTI, and ≥ 1 inpatient visit in the past 2 years) had a substantially elevated relative risk of 20.44 and AF_e of 0.93; however, such patients represented only 0.7% of the study population (n = 152,661).

Discussion

The study has demonstrated the use of machine learning models and large real-world datasets to identify correlating features and potential risk factors for infectious diseases, such as IED. By using a data-driven approach that required minimal clinical input, the model identified a variety of features that correlated with increased odds of a patient developing IED within 2 years.

In our model, SHAP values were calculated for all features and patients to assess the contribution of each feature to the model's predictions. Dependence plots of the SHAP values and feature values were used to further understand the relationship between a feature and the risk of the outcome, as encoded by the model. Although such dependencies do not necessarily represent causal relationships, they may represent clinically meaningful phenomena that warrant further investigation.

By interpreting these complex results as binary risk-exposure variables (e.g., "age ≥ 60" or "any prior history of UTI"), the impact of these features could be expressed in common epidemiologic metrics, such as relative risk and attributable risk fraction, enabling comparisons to be made with results from other models, data sources, and clinical literature. Moreover, features could be analyzed jointly to understand whether they represent potentially independent sources of risk. This approach examines only correlations between these factors and IED in the study population and does not identify causal relationships, but the transformation of these complex features into simple binary terms and the resulting analysis permit future external validation and investigation.

The findings of this model are concordant with the literature. Previous studies have established that the incidence of *E. coli* bacteremia increases with age [13–15]. A descriptive epidemiology study conducted in England revealed that 70.5% of ExPEC bacteremia cases occurred in patients aged ≥ 65 years [15], with the highest rate in men aged ≥ 85 years. In this study, the primary focus of infection was most commonly related to the urinary, hepatobiliary, and gastrointestinal tracts, and approximately half of the cases of community-onset bacteremia were related to a history of undergoing healthcare interventions. A population-based cohort study involving administrative databases from a health maintenance organization described a higher incidence of *E. coli* bacteremia among patients ≥ 85 years than those aged 65 to 69 years [14]. Furthermore, a systematic literature review of 210 studies reported increasing incidence rates of *E. coli* bacteremia as patients' age increased from 60 to ≥ 80 years [13].

Additionally, our model indicated that a history of UTI increased the probability of IED, with more frequent or more recent UTIs further increasing the risk of developing IED. Previous studies have reported UTI as a source of IED, with indwelling urinary catheters, renal dialysis, and kidney transplantation also established as risk factors for *E. coli* bacteremia [13, 14]. Additional work is required to better understand the role of previous UTIs as an independent risk factor for IED.

The following risk factors for *E. coli* bacteremia have been previously identified in patients aged ≥ 65 years: indwelling urinary catheter in men (odds ratio [OR], 77.4; 95% confidence interval [CI], 9.50–630.33; $P < 0.001$); urinary incontinence without catheterization in men (OR, 6.78; 95% CI, 2.43–18.97; $P < 0.001$) and women (OR, 2.85; 95% CI, 1.51–5.38; $P < 0.001$); and chronic renal failure in women (OR, 25.72; 95% CI, 2.49–264.80; $P = 0.006$) [14]. Renal dialysis (relative risk, 26.9) and solid-organ transplantation (relative risk, 20.3) increased the risk of developing *E. coli* bacteremia, which represented 23–55% of all bacteremia cases following kidney transplantation [13].

Our model identified a history of hospitalization as a potential novel risk factor for developing IED. The link between an increasing number of ED and inpatient visits and a higher probability of IED suggests that a history of hospitalization may be an important factor for developing IED. These findings may suggest that patients who interact more often with the inpatient healthcare system show an increased propensity to develop IED, possibly reflecting an overall poorer health status and increased frailty. Although previous studies explored the differences between community- and hospital-acquired *E. coli* bacteremia cases, none discussed prior hospitalization as a potential risk factor for IED [13–15]. Further work needs to be done to understand whether the reason(s) for the prior hospitalization(s) and/or ED visits correlates with a differential risk of developing IED. The increased risk for IED identified by our model for patients with no recent outpatient visits suggests that having less access to (primary) healthcare and/or reduced medical monitoring may also contribute to a patient's risk of developing IED, which may be important in low-income or resource-poor settings.

In addition to univariate potential risk factors, our model and approach also identified combinations of potential risk factors that put patients at much higher risk for IED. Patients aged ≥ 60 years with a history of UTI were shown to be at substantially greater risk than either individual risk group alone. Patients who also had an inpatient visit in the past 2 years were at even higher risk. A similar analysis also revealed that the risk contributed by T2D, an important potential risk factor when considered in isolation, is strongly linked to patient age.

The model identified other features that correlated with IED, including high platelet distribution width and blood urea nitrogen, the presence of albumin in urine, and a recent increase in heparin. These potential predictive features for IED have not been described previously in the literature but likely reflect the existence of underlying medical conditions that are risk factors themselves. For example, high levels of blood urea nitrogen and the presence of albumin in urine may be indicative of chronic kidney disease, a known risk factor of developing IED. High platelet distribution width may predict mortality in patients with sepsis, based on previous studies [29–31].

For patients aged ≥ 60 years with a history of UTI, diabetes mellitus and history of kidney disease have been identified as risk factors for IED in previous studies [12–14, 32]. According to a systematic review of 210 studies, the risk of developing *E. coli* bacteremia was increased in patients undergoing renal dialysis or solid-organ transplantation (with signs of kidney disease) [13]. Our study also showed that a history of kidney disease increases the probability of IED.

A population-based cohort study based on health maintenance organization data revealed higher rates of *E. coli* bacteremia among women with diabetes mellitus and men aged 80 to 84 years with diabetes mellitus [14]. Two studies, 1 from France and 1 from Brazil, reported that almost 20% of patients with *E. coli* bacteremia had diabetes mellitus [12, 32]. In our study, the prevalence of diabetes was approximately 30% among patients aged ≥ 60 years with a history of UTI; T2D increased the risk for predicting IED twofold.

Our model and approach to data interpretation have proved useful for initial screening and identification of potential risk factors and predictive features of IED. However, some limitations must be acknowledged. Most significantly, the model identifies only correlations between features and outcomes and does not necessarily establish any causal or clinical relationship between a feature and the outcome. Such relationships can be identified or confirmed only through causal analysis, clinical validation, and potentially the use of randomized, controlled trials.

Without such corroboration, it is possible some of these correlations are spurious or due to technical artifacts of the data.

Additionally, the model is limited by the fact that it is intrinsically retrospective and conducted on non-research–grade Optum® EHR data. Specifically, the model is sensitive to how the outcome is defined in the database and any biases inherent in the data (including those resulting from inconsistent collection or data missing not at random). For instance, previous studies identified that certain specific procedures and consumption of certain foods may increase the risk for IED [13, 33–35]. In this study, the dataset may not capture all procedures and patient lifestyle choices.

Moreover, the model cannot identify high-risk sequences of events. As described, the temporal relationship between occurrences of an event in a patient's medical record is compressed to a scalar value for each event. This makes the current approach infeasible for identifying sequences of different events that together may identify a uniquely high-risk individual, or even for identifying exactly how the pattern of a single event may contribute to the outcome.

Considering these limitations, we propose that our approach could be used as a preliminary screening tool to identify potential risk factors for a specific disease or outcome with minimal pre-specification or domain knowledge. Subsequently, the identified risk factors may be evaluated and/or confirmed using literature searches or traditional risk factor identification approaches (e.g., logistic regression).

Conclusions

This study represents an advancement in the scale of data for which this model may be employed for infectious disease research. Earlier infectious disease studies have applied XGBoost models to data from smaller population samples, ranging from 100 to 4000 patients in up to 33 hospitals [18–21]. Our analysis included more than 22 million patients from > 700 hospitals and > 7000 clinics.

The identified potential risk factors for IED must now be investigated in independent cohorts using more conventional methods. By determining the prevalence of these risk factors in the population, we will be able to confirm their usefulness in predicting IED, and they can then be used to improve our understanding of the disease and aid in treatment and prevention strategies.

Abbreviations

2-D	Two-dimensional
AF_e	Attributable fraction among the exposed
CI	Confidence interval
E. Coli	*Escherichia coli*
ED	Emergency department
HER	Electronic health record
ExPEC	Extraintestinal pathogenic Escherichia coli
IED	Invasive Escherichia coli disease
OR	Odds ratio
SHAP	Shapley Additive exPlanation
T2D	Type 2 diabetes
UTI	Urinary tract infection

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-024-09669-3.

Supplementary Material 1.

Acknowledgements

Medical writing support was provided by Janna Afanasjeva and Philip Matthews of Eloquent Scientific Solutions.

Data sharing plans

Data, documentation, and code used in this analysis will be available upon reasonable request from Erik Clarke, eclarke@its.jnj.com.

Authors' contributions

EC, CC, NK, BS, JP, and JG contributed to the design of this study. EC performed data analysis, and all authors contributed to the interpretation of the data. EC and JG drafted the manuscript, and CC, NK, BS, JP, and JG reviewed and revised drafts for scientific content. All authors gave their final approval of the draft and are accountable for its accuracy.

Funding

This study was sponsored by Janssen Research & Development, LLC, and all authors are employed by them.

Declarations

Ethics approval and consent to participate

Patient-level data extracted from the Optum® de-identified Electronic Health Record dataset were anonymous as per the US government's Health Insurance Portability and Accountability Act (HIPAA) guidelines. The HIPAA Privacy Rule ensures the protection of individuals' health information while allowing the necessary exchange of health information to inform medical research and promote high quality health care. Optum's de-identification is under expert determination and there are no restrictions on the use or disclosure of de-identified health information. For more information on the HIPAA Privacy Rule, please visit https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/. This study was conducted in compliance with the ethical principles of the Declaration of Helsinki and consistent with good clinical practice.

Consent for publication

The authors and Optum Pan-Therapeutic allow the editor to publish this work.

Competing interests

EC, CC, NK, BS, JP, and JG are employees of Janssen.

Received: 7 November 2023 Accepted: 25 July 2024
Published online: 08 August 2024

References

1. Braz VS, Melchior K, Moreira CG. *Escherichia coli* as a multifaceted pathogenic and versatile bacterium. Front Cell Infect Microbiol. 2020;10:548492.
2. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. Global extraintestinal pathogenic *Escherichia coli* (ExPEC) lineages. Clin Microbiol Rev. 2019;32(3):e00135-e218.

3. Russo TA, Johnson JR. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of *Escherichia coli*: ExPEC. J Infect Dis. 2000;181(5):1753–4.
4. Johnson JR, Russo TA. Extraintestinal pathogenic *Escherichia coli*: "the other bad E coli." J Lab Clin Med. 2002;139(3):155–62.
5. Owrangi B, Masters N, Kuballa A, O'Dea C, Vollmerhausen TL, Katouli M. Invasion and translocation of uropathogenic *Escherichia coli* isolated from urosepsis and patients with community-acquired urinary tract infection. Eur J Clin Microbiol Infect Dis. 2018;37(5):833–9.
6. Sarowska J, Futoma-Koloch B, Jama-Kmiecik A, Frej-Madrzak M, Ksiazczyk M, Bugla-Ploskonska G, et al. Virulence factors, prevalence and potential transmission of extraintestinal pathogenic *Escherichia coli* isolated from different sources: recent reports. Gut Pathog. 2019;11:10.
7. Pitout JD. Extraintestinal pathogenic *Escherichia coli*: a combination of virulence with antibiotic resistance. Front Microbiol. 2012;3:9.
8. Santos AC, Zidko AC, Pignatari AC, Silva RM. Assessing the diversity of the virulence potential of *Escherichia coli* isolated from bacteremia in São Paulo. Brazil Braz J Med Biol Res. 2013;46(11):968–73.
9. Johnson JR, Murray AC, Gajewski A, Sullivan M, Snippes P, Kuskowski MA, et al. Isolation and molecular characterization of nalidixic acid-resistant extraintestinal pathogenic *Escherichia coli* from retail chicken products. Antimicrob Agents Chemother. 2003;47(7):2161–8.
10. Laupland KB, Church DL. Population-based epidemiology and microbiology of community-onset bloodstream infections. Clin Microbiol Rev. 2014;27(4):647–64.
11. Poolman JT, Wacker M. Extraintestinal pathogenic *Escherichia coli*, a common human pathogen: challenges for vaccine development and progress in the field. J Infect Dis. 2016;213(1):6–13.
12. Daga AP, Koga VL, Soncini JGM, de Matos CM, Perugini MRE, Pelisson M, et al. *Escherichia coli* bloodstream infections in patients at a university hospital: virulence factors and clinical characteristics. Front Cell Infect Microbiol. 2019;9:191.
13. Bonten M, Johnson JR, van den Biggelaar AHJ, Georgalis L, Geurtsen J, de Palacios PI, et al. Epidemiology of *Escherichia coli* bacteremia: a systematic literature review. Clin Infect Dis. 2021;72(7):1211–9.
14. Jackson LA, Benson P, Neuzil KM, Grandjean M, Marino JL. Burden of community-onset *Escherichia coli* bacteremia in seniors. J Infect Dis. 2005;191(9):1523–9.
15. Bou-Antoun S, Davies J, Guy R, Johnson AP, Sheridan EA, Hope RJ. Descriptive epidemiology of *Escherichia coli* bacteraemia in England, April 2012 to March 2014. Euro Surveill. 2016;21(35):30329.
16. Chen T, Guestrin C, editors. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016. https://doi.org/10.1145/2939672.2939785.
17. Sang S, Sun R, Coquet J, Carmichael H, Seto T, Hernandez-Boussard T. Learning From Past Respiratory Infections to Predict COVID-19 Outcomes: Retrospective Study. J Med Internet Res. 2021;23(2):e23026.
18. Guan X, Zhang B, Fu M, Li M, Yuan X, Zhu Y, et al. Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. Ann Med. 2021;53(1):257–66.
19. Bertsimas D, Lukin G, Mingardi L, Nohadani O, Orfanoudaki A, Stellato B, et al. COVID-19 mortality risk assessment: an international multi-center study. PLoS ONE. 2020;15(12):e0243262.
20. Liu J, Zhang S, Dong X, Li Z, Xu Q, Feng H, et al. Corticosteroid treatment in severe COVID-19 patients with acute respiratory distress syndrome. J Clin Invest. 2020;130(12):6417–28.
21. Garcia-Vidal C, Puerta-Alcalde P, Cardozo C, Orellana MA, Besanson G, Lagunas J, et al. Machine learning to assess the risk of multidrug-resistant gram-negative bacilli infections in febrile neutropenic hematological patients. Infect Dis Ther. 2021;10(2):971–83.
22. Optum® de-identified Electronic Health Record dataset (2007–2020).
23. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017. In: Advances in neural information processing systems 30. Curran Associates, Inc.; 4765–74. Available from: https://ui.adsabs.harvard.edu/abs/2017arXiv170507874L.
24. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969–75.
25. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54–60.
26. Medical Dictionary for Regulatory Activities Herndon, VA, USA: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use; 2022. Available from: https://www.meddra.org/.
27. First Databank Knowledge® Foundations South San Francisco, CA, USA: First Databank, Inc.; 2022. Available from: https://www.fdbhealth.com/solutions/medknowledge-drug-database/medknowledge-foundations.
28. Thorwarth WT Jr. CPT: an open system that describes all that you do. J Am Coll Radiol. 2008;5(4):555–60.
29. Guclu E, Durmaz Y, Karabay O. Effect of severe sepsis on platelet count and their indices. Afr Health Sci. 2013;13(2):333–8.
30. Orak M, Karakoç Y, Ustundag M, Yildirim Y, Celen MK, Güloglu C. An investigation of the effects of the mean platelet volume, platelet distribution width, platelet/lymphocyte ratio, and platelet counts on mortality in patents with sepsis who applied to the emergency department. Niger J Clin Pract. 2018;21(5):667–71.
31. Mangalesh S, Dudani S, Malik A. Platelet indices and their kinetics predict mortality in patients of sepsis. Indian J Hematol Blood Transfus. 2021;37:600–8.
32. Lefort A, Panhard X, Clermont O, Woerther PL, Branger C, Mentré F, et al. Host factors and portal of entry outweigh bacterial determinants to predict the severity of *Escherichia coli* bacteremia. J Clin Microbiol. 2011;49(3):777–83.
33. Fibke CD, Croxen MA, Geum HM, Glass M, Wong E, Avery BP, et al. Genomic epidemiology of major extraintestinal pathogenic *Escherichia coli* lineages causing urinary tract infections in young women across Canada. Open Forum Infect Dis. 2019;6(11):ofz431.
34. Chen YC, Chang CC, Chiu THT, Lin MN, Lin CL. The risk of urinary tract infection in vegetarians and non-vegetarians: a prospective study. Sci Rep. 2020;10(1):906.
35. Rosenberg S, Bonten M, Haazen W, Spiessens B, Abbanat D, Go O, et al. Epidemiology and O-serotypes of extraintestinal pathogenic *Escherichia coli* disease in patients undergoing transrectal ultrasound prostate biopsy: a prospective multicenter study. J Urol. 2021;205(3):826–32.

Publisher's Note