Annals of Medicine and Surgery 84 (2022) 104956

Contents lists available at ScienceDirect

Annals of Medicine and Surgery

journal homepage: www.elsevier.com/locate/amsu

Systematic Review / Meta-analysis

Performance of machine learning algorithms for surgical site infection case detection and prediction: A systematic review and meta-analysis

Guosong Wu [a,b,e,*], Shahreen Khair [a], Fengjuan Yang [a], Cheligeer Cheligeer [c], Danielle Southern [a,b], Zilong Zhang [a], Yuanchao Feng [a], Yuan Xu [a,d], Hude Quan [a,b], Tyler Williamson [a,b], Cathy A. Eastwood [a]

[a] Centre for Health Informatics, Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[b] O'Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[c] Alberta Health Services, Calgary, Alberta, Canada
[d] Department of Oncology and Surgery, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[e] Institute of Health Economics, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

Keywords:
Surgical wound infection
Machine learning
Algorithms
Systematic review
Meta-analysis

ABSTRACT

Background: Medical researchers and clinicians have shown much interest in developing machine learning (ML) algorithms to detect/predict surgical site infections (SSIs). However, little is known about the overall performance of ML algorithms in predicting SSIs and how to improve the algorithm's robustness. We conducted a systematic review and meta-analysis to summarize the performance of ML algorithms in SSIs case detection and prediction and to describe the impact of using unstructured and textual data in the development of ML algorithms.

Methods: MEDLINE, EMBASE, CINAHL, CENTRAL and Web of Science were searched from inception to March 25, 2021. Study characteristics and algorithm development information were extracted. Performance statistics (e.g., sensitivity, area under the receiver operating characteristic curve [AUC]) were pooled using a random effect model. Stratified analysis was applied to different study characteristic levels. Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Diagnostic Test Accuracy Studies (PRISMA-DTA) was followed.

Results: Of 945 articles identified, 108 algorithms from 32 articles were included in this review. The overall pooled estimate of the SSI incidence rate was 3.67%, 95% CI: 3.58–3.76. Mixed-use of structured and textual data-based algorithms (pooled estimates of sensitivity 0.83, 95% CI: 0.78–0.87, specificity 0.92, 95% CI: 0.86–0.95, AUC 0.92, 95% CI: 0.89–0.94) outperformed algorithms solely based on structured data (sensitivity 0.56, 95% CI:0.43–0.69, specificity 0.95, 95% CI:0.91–0.97, AUC = 0.90, 95% CI: 0.87–0.92).

Conclusions: ML algorithms developed with structured and textual data provided optimal performance. External validation of ML algorithms is needed to translate current knowledge into clinical practice.

1. Introduction

Surgical site infections (SSIs) are the most frequently reported healthcare-associated infections among surgical patients [1,2]. Annually, 1.3 million operative procedures are performed in Canada, and an estimated 312.9 million operations are completed globally, of which 2–5% of the patients acquire SSIs [2,3]. The extra cost attributable to SSIs is estimated to be $20,842 USD per admission, and patient hospital stay is prolonged by an average of 9.7 days [4,5]. Most importantly, the rate of SSIs is increasing because of surgery volume growth and longer life expectancy [5,6].

Detecting SSIs is essential for infection prevention and control programs to further quality initiatives and decrease infection rates. Traditional SSI case identification methods rely on International Classification of Diseases 10th Revision (ICD-10) codes or chart reviews. However, the validity of ICD-10 codes in administrative databases varies

* Corresponding author. Centre for Health informatics and O'Brien Institute for Public Health, Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada.
E-mail address: guosong.wu@ucalgary.ca (G. Wu).

https://doi.org/10.1016/j.amsu.2022.104956
Received 13 September 2022; Received in revised form 8 November 2022; Accepted 13 November 2022
Available online 23 November 2022

by local coding practices, and chart review requires extensive human resources and is often time-consuming [7]. Machine learning (ML) algorithms that leverage rich text data documented in electronic medical records (EMR) have been applied in SSI case identification and prediction [8–12]. However, to date, the effectiveness of these ML algorithms has not been summarized.

To close this knowledge gap, this systematic review aims to summarize literature evidence of ML algorithms' performance in SSIs case detection and prediction, describe the impact of the use of unstructured and textual data in the performance of ML algorithms, and provide summaries of methodologies commonly applied in ML algorithm development.

2. Methods

The review protocol was registered at PROSPERO (register number: CRD42022339630) [13]. The Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Diagnostic Test Accuracy Studies (PRISMA-DTA) guidelines were followed [14].

2.1. Search strategy

The literature search strategy (Supplement, eAppendix) was developed with the help of an experienced medical information expert. The Ovid MEDLINE, Ovid EMBASE, CINAHL, Cochrane Central Register of Controlled Trials (CENTRAL) and Web of Science databases were systematically searched from inception to March 25, 2021. Two groups of subject headings and synonyms of machine learning and surgical site infection were combined and searched without language limitation. Non-human studies, case reports, editorials, protocols, comments, or letters were excluded (Supplement: Search Strategy). Grey literature search was conducted through OpenGrey and Google Scholar.

2.2. Eligibility criteria

Only original studies were included. An article was considered for inclusion if it met the following criteria: (1) Population: adult patients who underwent any type of surgery (e.g., colorectal, gastrointestinal, orthopedic, abdominal, neurosurgical, etc.). (2) Measures: Authors defined ML algorithms for detecting and/or predicting SSIs. These include but are not limited to LASSO model, decision trees, random forest, artificial neural network, etc. (3) Comparison: Reference standard in the article to confirm the presence of SSIs. (4) Outcomes: The performance measures of ML algorithms. (5) Study Design: Not limited.

.

2.3. Study selection

During the initial round of title and abstract screening, two reviewers (GW and SK), independently and in duplicate, reviewed the titles and abstracts for all retrieved citations. The same two reviewers subsequently reviewed the full texts of abstracts identified by both reviewers during the first screen. Articles that met the above inclusion criteria were included in the data extraction. Kappa statistic was employed in both the screening stages to measure agreement between reviewers [15]. All citations were managed with EndNote 20 (Thomas Reuters, Philadelphia, PA, USA).

2.4. Data extraction

The information of each selected article was collected in a data extraction form developed prior to review. The two reviewers extracted the following data independently: study characteristics (e.g. publication year/country, funding source, study design, etc.), patient demographic information (age, sex), SSIs information (e.g., type of wound, type of SSIs, the incidence rate of SSIs, etc.), ML algorithm information (e.g., data source/sample size of model training and validation) and performance measures (i.e., sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV], accuracy and area under the receiver operating characteristic curve [AUC]). The reviewers (GW and SK) extracted the numbers of true positive, false positive, false negative and true negative of SSIs (two-by-two table) for each algorithm, if reported in the article, or recalculated them (using available data) with the online Diagnostic Test Calculator, if not reported [16].

2.5. Risk of bias assessment

Two reviewers (GW, FY) independently assessed the risk of bias for including articles using the validated tool QUADAS 2 (Quality Assessment tool for Diagnostic Accuracy Studies) [17]. Rating results were discussed, and a consensus was reached. RevMan 5 was used to generate the risk of bias and applicability concerns summary and graph (Review Manager (RevMan) [Computer program]. Version 5.4, The Cochrane Collaboration, 2020).

2.6. Synthesis of results

Discrepancies from any review procedures were resolved by consensus. A third reviewer (DS) was involved when necessary. A PRISMA flow diagram was applied to indicate the number of articles included or excluded in the review and meta-analysis. Descriptive statistics were calculated for the results from the extracted data, including study characteristics, the incidence rate of SSIs, and ML performance indicators. Meta-analysis was performed to examine the performance estimates of selected ML algorithms along with a confusion matrix and its 95% confidence interval (CI) under a random-effects model. Stratified analysis was applied to explore performance at different levels.

The source of heterogeneity in a systematic review of DTA studies includes within-study variabilities (among ML algorithms) and between-study differences [18–20]. Therefore, heterogeneity was presumed in this review. Traditional heterogeneity measurements (e.g., Cochrane Q and I^2) were univariate tests that do not account for heterogeneity among different ML algorithms within each study [18]. We followed the Cochrane Handbook for Systematic Reviews of DTA studies to graphically depict the observed heterogeneity using the summary receiver operating characteristic (SROC) curve [18,21]. The test performance was estimated with the hierarchical summary receiver operating characteristic (HSROC) model. The model utilizes a hierarchical structure of data distributions in terms of two levels, within-study variability, and between-study variability [18–20]. It can provide equivalent summary estimates for both sensitivity and specificity. The overall ML algorithm performance was pooled using HSROC model, median and its corresponding interquartile ranges (IQR) were also presented [19,20]. The Deeks Funnel Plot was used to determine publication bias [22]. Likelihood Ratio Scatter was applied to graphically display the potential applications of developed algorithms [23]. All statistical analyses were performed using Stata SE 16 (StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC.).

3. Results

The literature search initially revealed 945 articles. Title and abstract review resulted in 59 articles for full-text review (Kappa = 0.701), 17 articles from a review of references and six articles from grey literature search. A total of 32 articles [7,8,10,24–52] were included for qualitative synthesis with a Kappa agreement of 0.805. We conducted a meta-analysis on 15 articles [7,10,40–52] with cohort study design and sufficient ML performance data to enable calculations (Fig. 1).

3.1. Characteristics of included articles

As summarized in Table 1, over half of the included studies were

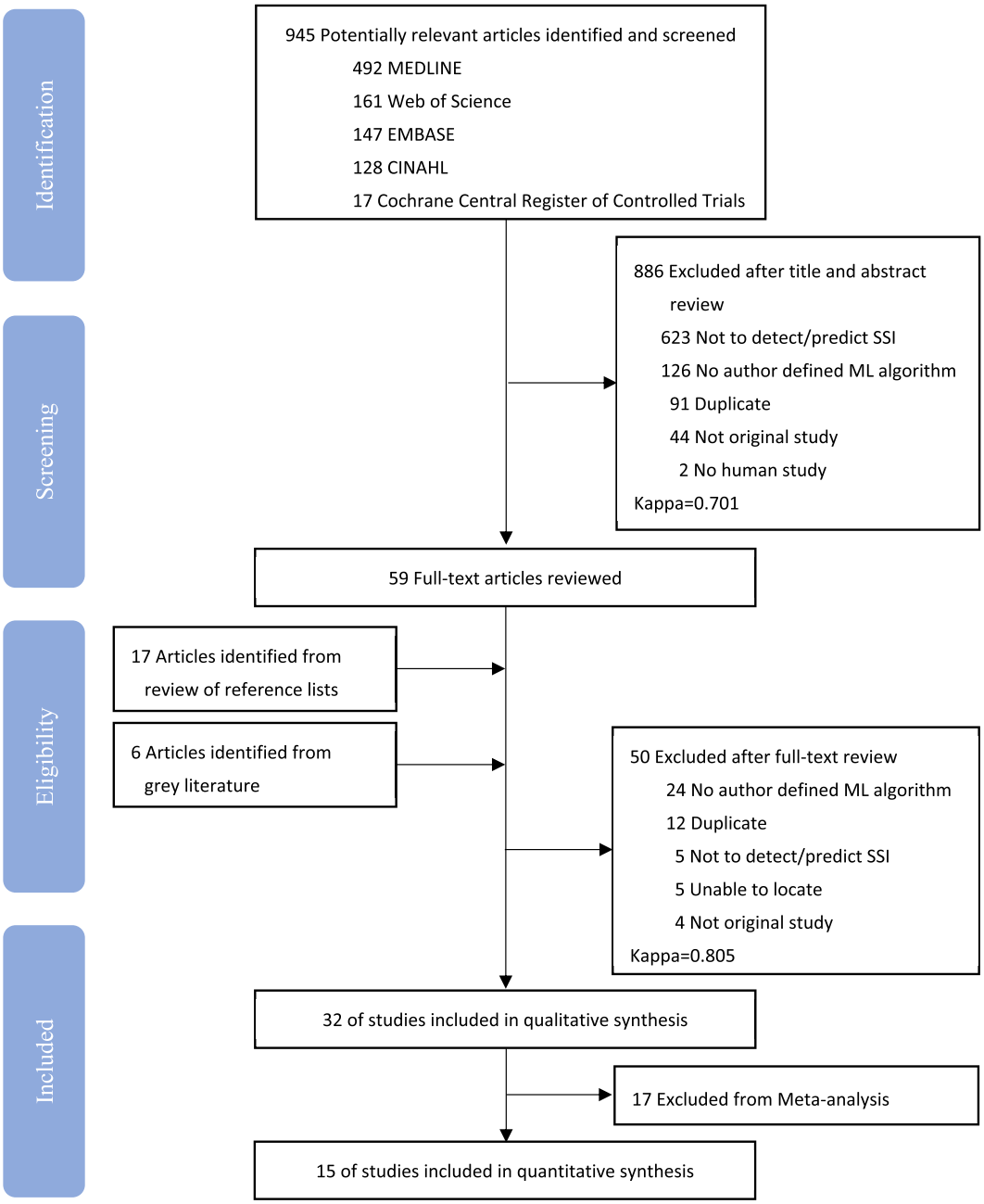

Fig. 1. Selection of articles for Review.
Abbreviations: SSI, surgical site infections; ML, machine learning.

published between 2019 and 2020. Most of them were retrospective (81.3%), single-center (65.6%) studies originating in the United States (68.8%) using a consecutive sampling method (56.3%), with a focus on mixed (37.5%), colorectal (15.6%) or gastrointestinal (15.6%) surgical procedures. Of the 32 articles, 23 (71.9%) reported SSI follow-up days, of which 17 (73.9%) reported 30 days post-surgery follow-up. The majority (56.3%) of the studies did not disclose any funding source.

Wound classification (clean 49.2%, clean/contaminated 27.2%, contaminated 15.3% and dirty 8.3%) was reported in three articles [10,33,48]. Excluding studies that did not report on sample sizes for model training and/or testing [24,25,31,39] and case control studies [30,35–37], a total of 6076 SSIs occurred in 165,717 surgical procedures. This translated to an overall pooled estimate of SSIs incidence rate of 3.67%, 95% CI: 3.58–3.76. Types of SSIs were categorized as superficial incisional SSI (n = 1,074, 42.84%), deep incisional SSI (n = 427, 17.03), and organ/space SSI (n = 1,006, 40.13%) in five included studies [10,27,28,43,52].

3.2. Risk of bias in articles

Four articles failed to report on the detailed process of patient selection [30,35–37,53]. Reference standard used for algorithm validation was not clearly stated in nine articles [10,25,29,30,33,36,38,43,48]. The applicability concerns were minimal among all included articles (Supplement, eFig. 1).

3.3. Description of ML algorithms

A total of 108 ML algorithms were retrieved from the included studies, 36 (33.3%) were detective algorithms, and 72 (66.7%) were

Table 1
Characteristics of included articles.

Study ID	Country	Population	Study Design	Setting	Sampling Method	Patient Age	Female No. (%)	Surgery Type	SSI Follow-up Days	Funding Source
Atti/2020 [43]	Italy	Pediatric	Retro	Single	Consecutive	NR	NR	Mixed	30	NR
Azimi/2020 [46]	USA	NR	Retro	Single	Consecutive	Mean = 60.25	106 (51.5)	Colorectal	NR	NR
Bucher/2020 [10]	USA	NR	Retro	Multiple	Random	Mean = 53, SD = 18	11,077 (50.8)	Mixed	30	G
Chen/2020 [39]	China	Adult	Retro	Multiple	Consecutive	Median = 54.3, IQR (44, 65)	13,293 (61.5)	Mixed	NR	NR
Hopkins/2020 [44]	USA	NR	Retro	Single	Consecutive	57.5	2104 [52]	Orthopedic	NR	NR
Karhade/2020 [38]	USA	Adult	Retro	Multiple	Consecutive	Median = 47, IQR [37,59]	2628 (44.8)	Orthopedic	90	P
Merath/2020 [29]	USA	Adult	Retro	Multiple	Consecutive	Median = 66, IQR (57, 75)	NR	Colorectal	30	NR
Petrosyan/2020 [47]	Canada	Adult	Retro	Single	Consecutive	Mean = 56.7	4540 (56.3)	Mixed	30	G
Skube/2020 [50]	USA	NR	Retro	Single	Consecutive	Mean = 54, SD = 17	5094 [45]	Mixed	NR	G
Song/2020 [49]	USA	Mixed	Retro	Multiple	Consecutive	NR	3990 [43]	Cardiac and vascular	NR	G
Gowd/2019 [23]	USA	NR	Retro	Multiple	Consecutive	Mean = 69.5, SD = 9.6	7493 (43.8)	Mixed	30	NR
Quérouéa/2019 [48]	France	NR	Retro	Single	Consecutive	NR	NR	Orthopedic	90	NR
Shen/2019 [32]	USA	NR	Retro	Single	NR	NR	NR	Colorectal	30	G
Shi/2019 [40]	USA	NR	Retro	Multiple	NR	NR	NR	Mixed	30	G
Silva/2019 [7]	Brazil	Pediatric	Retro	Single	NR	Mean = 48.31, SD = 22.03	7107 (56.9)	Mixed	30	NR
Thirukumaran/2019 [35]	USA	NR	CC	Single	Consecutive	Mean = 45.8, SD = 20.7	760 [48]	Orthopedic	90	G
Tunthanathip/2019 [41]	Thailand	NR	Retro	Single	Random	Mean = 45.1, SD = 21.1	632 [43]	Neurosurgical	90	NR
Grundmeier/2018 [24]	USA	Pediatric	Retro	Multiple	Consecutive	NR	NR	Mixed	60	G
Kocbek/2018 [42]	Slovenia	NR	Retro	Single	NR	NR	NR	Gastrointestinal	60	G
Strauman/2018 [34]	Norway	NR	CC	Single	NR	NR	NR	Gastrointestinal	NR	G
Weller/2018 [36]	USA	NR	Retro	Single	Random	NR	NR	Colorectal	NR	NR
Chapman/2017 [37]	USA	NR	Retro	Multiple	NR	NR	NR	Gastrointestinal	30	G
Sohn/2017 [8]	USA	NR	Retro	Single	Random	NR	NR	Colorectal	30	NR
Hu/2016 [26]	USA	Pediatric	Retro	Single	Consecutive	NR	NR	Mixed	30	NR
Ke/2016 [27]	USA	NR	Pro	Single	NR	Mean = 56.4	345 (63.7)	General-abdominal	30	NR
Mandagani/2016 [28]	USA	NR	CC	NR	NR	NR	NR	Gastrointestinal	NR	NR
Sanger/2016 [31]	USA	NR	Pro	Single	Random	Mean = 56.5	542 (63.7)	General-abdominal	30	NR
Hu/2015 [25]	USA	NR	Retro	Single	Consecutive	NR	3754 [60]	Mixed	30	G
Soguero-Ruiz/2015 [33]	Norway	NR	CC	Single	Random	Mean = 57.0, SD = 20.7	477 (47.4)	Gastrointestinal	NR	NR
Esbroeck/2014 [22]	USA	NR	Retro	Multiple	Consecutive	NR	NR	Mixed	30	NR
Michelson/2014 [30]	USA	Adult	Retro	Single	Consecutive	Mean = 53.5	1058 (48.4)	Mixed	30	G
Campillo-Gimenez/2013 [45]	France	Adult	Retro	Single	Consecutive	NR	NR	Neurosurgical	30	NR

Abbreviation: CI, confidence interval; IQR, interquartile range; NR, not reported; SD, standard deviation; SSI, surgical site infection.
Study Design: Retro, retrospective cohort study; Pro, prospective cohort study; CC, case control study; **Setting**: Single, single center; Multiple, multiple centers; **Funding source**: G, government; P, private.

predictive algorithms. There were nine (8.3%) algorithms developed from textual data only, 47 (43.5%) algorithms developed from structured data and 52 (48.2%) algorithms developed with a mixed data source. The reference standard data sets used for ML algorithm development were from national surveillance or quality improvement programs (n = 8, 25%), hospital surveillance programs (n = 5, 15%) and chart review (n = 8, 25%). The ML algorithms were developed using Logistic Regression and its variation (n = 31, 28.7%), random forest (n = 13, 12%), decision tree (n = 8, 7.4%), support vector machines (n = 8, 7.4%), and Bayesian network (n = 7, 6.5%). The model threshold was reported in three (9.4%) articles [40,44,52]. The median sample size and SSI cases were 3410 (IQR: 1115–5992) and 167 (IQR: 60–232) for model training, 1616 (IQR: 654-4160) and 142 (IQR: 34–216) for model validation, respectively (Table 2).

3.4. Performance of ML algorithms

The research team extracted the two-by-two table for 15 articles that included 44 algorithms (7 detective and 37 predictive) [7,10,40–52]. The ML algorithms' median sensitivity and specificity were 0.78 (IQR: 0.62–0.86) and 0.91 (IQR: 0.87–0.98), respectively. HSROC model indicated an overall pooled sensitivity of 0.74, 95% CI: 0.66–0.81, specificity of 0.95, 95% CI 0.92–0.97, and AUC of 0.93, 95% CI: 0.90–0.95. Heterogeneity was depicted with SROC cure and explored with stratified analysis. Most of the observed study results lay close to the summary ROC curve (Supplement, eFig. 2); however, there were quite a few scattered in ROC space, indicating a certain amount of heterogeneity. We explored the heterogeneity and its source of origin with meta-regression adjusted for study characteristics [54,55]. The type of ML algorithms (Predictive/Detective), mixed-use of structured

Table 2
Summaries of machine learning algorithms development.

Study ID	Data		ML Technique	Model Type	Model Threshold	Model Training (REF)[a]			Model Validation (REF)		
	Type	Reference Standard				Sample Size	No. of SSI	SSI Incidence Rate (%)	Sample Size	No. of SSI	SSI Incidence Rate (%)
[b] Atti/2020 [43]	S-EMR, FT	Hospital surveillance program	RE	D	NR	T (2,944)	T [18]	T (0.61)	NR	NR	NA
[b] Azimi/2020 [46]	S-EMR	NR	BN, DT, SVM, ANNs, RF	P	NR	T (208)	T [18]	T (8.65)	NR	NR	NA
[b] Bucher/2020 [10]	S-EMR	NR	NER	D	NR	4574	255	5.57	17,210	793	4.61
[b] Chen/2020 [39]	S-EMR, FT	National surveillance data (NNIS)	LR, RF, DT, ANNs	P	NR	17,597	202	1.15	4014	43	1.07
[b] Hopkins/2020 [44]	NR	Chart review	ANNs	P	NR	3034	T [60]	T (1.48)	1012	NR	NA
[b] Karhade/2020 [38]	FT	Chart review	BC	D	0.05, 0.1, 0.5	4483	46	1.03	1377	16	1.16
Merath/2020 [29]	S-EMR	ACS-NSQIP	DT	P	NR	15,657	NR	NA	NR	NR	NA
[b] Petrosyan/2020 [47]	ADMIN	ACS-NSQIP	RF, LR	P	NR	10,046	556	5.53	4305	239	5.55
[b] Skube/2020 [50]	S-EMR	ACS-NSQIP	LR	D	0.04, 0.06	6188	398	6.43	5132	161	3.14
[b] Song/2020 [49]	ADMIN, S-EMR	National surveillance data (NIC-HAI)	LR, DT, SVM	P	NR	7419	T (205)	T (2.21)	1855	NR	NA
Gowd/2019 [23]	S-EMR	NR	LR	P	NR	13,697	NR	NA	3422	NR	NA
[b] Quérouéa/2019 [48]	S-EMR, FT	Hospital surveillance program	LR	D	NR	T (2,133)	T [22]	T (1.03)	NR	NR	NA
Shen/2019 [32]	FT	Chart review	DT, SVM, RF	D	NR	T (1,178)	T (80)	T (6.79)	NR	NR	NA
[b] Shi/2019 [40]	S-EMR, FT	Chart review	RF, SVM, LR	P	NR	T (5,795)	T (291)	T (5.02)	NR	NR	NA
[b] Silva/2019 [7]	S-EMR, FT	Hospital surveillance program	RF, LR, SVM, BN, NC, SGD	P/D	NR	15,479	188	1.21	12,637	202	1.60
Thirukumaran/2019 [35]	ADMIN, S-EMR, FT	Hospital surveillance program	LR	P	NR	1263	172	13.62	316	36	11.39
[b] Tunthanathip/2019 [41]	S-EMR	NR	DT, BN, ANNs, KNN	P	NR	T (1,471)	T (67)	T (4.55)	295	NR	NA
Grundmeier/2018 [24]	ADMIN, S-EMR, FT	Chart review	RF, LR	P	NR	6871	209	3.04	1039	25	2.41
[b] Kocbek/2018 [42]	S-EMR, FT	NR	LR, BC	P	Range: 0.171-0.245	909	183	20.13	228	50	21.93
Strauman/2018 [34]	S-EMR	ICD-10 and Procedure codes	ANNs	D	NR	T (883)	T (232)	T (26.27)	NR	232	NA
Weller/2018 [36]	S-EMR, FT	NR	LR, RF, SVM, BN, BC	P	NR	1051	102	9.71	232	18	7.76
Chapman/2017 [37]	S-EMR, FT	Chart review	SVM	D	NR	565	NR	NA	100	NR	NA
Sohn/2017 [8]	S-EMR, FT	Chart review	BN	D	NR	T (751)	T (67)	T (8.92)	NR	NR	NA
Hu/2016 [26]	S-EMR	ACS-NSQIP	LR	D	NR	5280	336	6.36	3629	157	4.33
Ke/2016 [27]	S-EMR, FT	NR	Linear regression, SVM	P	NR	652	T (167)	T (20.49)	163	NR	NA
Mandagani/2016 [28]	S-EMR	NR	LR, DT	P	NR	T (879)	T (181)	T (20.59)	NR	NR	NA
Sanger/2016 [31]	S-EMR	NR	BN	D	NR	T (851)	T (167)	T (19.62)	NR	229	NA
Hu/2015 [25]	S-EMR	ACS-NSQIP	LR	D	NR	3996	278	6.96	2262	127	5.61
Soguero-Ruiz/2015 [33]	S-EMR	ICD-10 and Procedure codes	SVM	P	NR	T [1,005]	T (101)	T (10.05)	NR	NR	NA
Esbroeck/2014 [22]	S-EMR, FT	ACS-NSQIP	LR	P	NR	602,089	NR	NA	350,545	NR	NA
Michelson/2014 [30]	S-EMR, FT	Hospital surveillance program	LR	P	NR	T (2,407)	T [59]	T (2.45)	NR	NR	NA
[b] Campillo-Gimenez/2013 [45]	S-EMR, FT	Chart review	VSM	D	NR	3785	42	1.11	1225	NR	NA

Abbreviation: IQR, interquartile range; NA, not available; NR, not reported; SSI, surgical site infection.
Data type: FT, Free text data; ADMIN, Administrative data; S-EMR, Structured Electronic Medical Records. **Reference standard**, ACS-NSQIP, American College of Surgeons-National Surgical Quality Improvement Program; NIC-HAI, Nursing Intensity of Patient Care Needs and Rates of Healthcare-Associated Infections; NNIS, National Nosocomial Infections Surveillance. **ML type**: ANNs, Artificial Neural Networks and its variations; BC; Boosted Classifiers (e.g., AdaBoost, XGBoost); BN, Bayesian Network; DT, Decision Tree; KNN, k-nearest neighbors; LR, Logistic Regression and its variations; NC, Nearest Centroid; NER, Named Entity Recognizer; SGD, Stochastic Gradient Descent; SVM, Support Vector Classification; RE, Regular Expression; RF, Random Forest; VSM, Vector Space Model. **Model type**: P, predictive; D, detective.
(REF): Data extracted from reference standard (e.g., chart review).
[a] T (number indicates total number of SSI/procedures (when training/testing sample size were not specified in article).
[b] Articles included for Meta-analysis.

and textual data sources for development of algorithms were associated with heterogeneity and were hence used for stratified analysis (Supplement, eFig. 3).

The median sensitivity and specificity of detective ML algorithms were 0.92 (IQR: 0.79–0.94) and 0.92 (IQR: 0.88–0.99), respectively. Pooled estimates in HSROC model reached a sensitivity of 0.89, 95% CI: 0.81–0.94, specificity of 0.98, 95% CI: 0.86–1.0 and AUC of 0.95, 95% CI: 0.93–0.97. The median sensitivity and specificity of predictive ML algorithms were 0.75 (IQR: 0.58–0.84) and 0.91 (IQR: 0.87–0.97), respectively. HSROC model pooled estimates had a sensitivity of 0.70, 95% CI: 0.61–0.78, specificity of 0.96, 95% CI: 0.91–0.96 and AUC of 0.92, 95% CI: 0.89–0.94. The median sensitivity and specificity of algorithms developed with structured data were 0.69 (IQR: 0.44–0.88) and 0.91 (IQR: 0.87–0.98), respectively. HSROC model pooled estimates had a sensitivity of 0.56, 95% CI:0.43–0.69, specificity of 0.95, 95% CI:0.91–0.97 and AUC of 0.90, 95% CI: 0.87–0.92. The median sensitivity and specificity of algorithms developed with mixed data sources were 0.84 (IQR: 0.79–0.89) and 0.90 (IQR: 0.87–0.92). HSROC model pooled estimates had a sensitivity of 0.83, 95% CI: 0.78–0.87, specificity of 0.92, 95% CI: 0.86–0.95, AUC of 0.92, 95% CI: 0.89–0.94 (Figs. 2 and 3).

The performance of different ML methodologies varied in this study. Using the HSROC model, Logistic Regression and its variation (n = 14) had a sensitivity of 0.82, 95% CI: 0.70–0.91, specificity of 0.91, 95% CI: 0.85–0.95, AUC = 0.94, 95% CI: 0.91–0.96, Artificial Neural Network (n = 5) had a sensitivity of 0.68, 95% CI: 0.50–0.82, specificity of 0.97, 95% CI: 0.88–0.99, AUC = 0.90, 95% CI:0.87–0.92, Random Forest (n = 5) had a sensitivity of 0.70, 95% CI: 0.56–0.81, specificity of 0.93, 95% CI: 0.86–0.96, AUC = 0.90, 95% CI: 0.87–0.93, Support Vector Machine (n = 4) had a sensitivity of 0.75, 95% CI: 0.56–0.87, specificity of 0.94, 95% CI: 0.86–0.96, AUC = 0.93, 95% CI:0.91–0.95.

3.5. Publication Bias Assessment

A mild asymmetric distribution of a natural logarithm of the DOR (x-axis) against a reciprocal of the square root of the effective sample size (y-axis) could be observed in Fig. 4. The results of Deeks funnel plot suggested symmetry (p = 0.14) of included studies and a low likelihood of publication bias.

3.6. Sensitivity analysis on robustness of study estimates

In this review, sensitivity analysis was undertaken by removing studies with application concerns rated by QUADAS 2, sample size smaller than one thousand, or keeping studies with 30 days of SSI follow-up. A repeat of the primary meta-analysis on algorithms derived from different data types was presented in eTable 1. In all sensitivity analyses, the ranking of the algorithms remained consistent with algorithms derived from mixed data sources outperforming algorithms solely developed from structured data in both measures of sensitivity and AUC. Furthermore, the magnitude and direction of differences remained similar across sensitivity analysis suggesting robust estimates.

3.7. Generalizability assessment

Most of the included studies (93.8%) were internally validated with k-fold cross-validation. However, only a single study evaluated the external validity of developed algorithms with a blind cohort from another healthcare system [10]. The potential application of developed ML algorithms was graphically summarized in Supplement eFig. 4. The solid red square in the scatter graph indicates the position of the combined positive likelihood ratio (LR) and negative LR estimates. The whiskers running through the red square are the confidence intervals for either positive LR (vertical whiskers) or negative LR (horizontal whiskers). The summary of the positive and negative likelihood ratios for ML algorithms with 95% CI in the upper right quadrant, indicates that the developed algorithms help confirm the presence of SSIs (when positive) and not their exclusion (when negative).

4. Discussion

Our review of the current literature identified 32 articles and 108 ML algorithms developed for SSI case detection and prediction. In addition, we observed an increased interest in applying ML techniques in SSI control and prevention with more articles published over the period studied. Despite a certain amount of heterogeneity, the median and IQR of raw data and HSROC model pooled estimates indicate that algorithms developed from mixed-use of structured data and textual data outperformed algorithms solely based on structured data. Among ML methods included in this review, Logistic Regression and its variation demonstrated superior performance, suggesting an important technique for future studies.

One important finding of this review is that adding clinical notes or free text as a data source for ML algorithm development could improve the model performance of SSIs case detection and prediction [26,41,56]. About 97% of SSIs occurred post-patient discharge [56]. Therefore, it is critical to detect or predict infections to capture any signs and symptoms documented in clinical notes with automated ML algorithms. Furthermore, our review suggests that ML algorithms trained with mixed-use of structured and textual data could produce comparable results compared with manual chart review. Current SSI surveillance programs mostly rely on ICD codes for an initial screen to exclude the most unlikely records, followed by a panel chart review to confirm the presence of SSI. The initial screen is crucial as the more accurate the rule-out methodology is, the fewer cases would remain for chart reviews, and subsequently be less time-consuming and cost-efficient. Given that the performance of ICD codes varied, it is anticipated that ML algorithms can be further developed and validated for SSI screen surveillance programs in the near future.

This meta-analysis revealed better performance for detective models compared to predictive models. This is not surprising given that detective algorithms were developed with hospitalization data, while most predictive algorithms were trained solely on data collected before surgery [7]. Logistic Regression and its variation were ranked at the top of ML algorithms that were included for meta-analysis and demonstrated its potential in automating SSI identification [57]. Depending on the purpose of a study, researchers need to choose the most relevant models (detective vs. predictive) and appropriate ML tools [58].

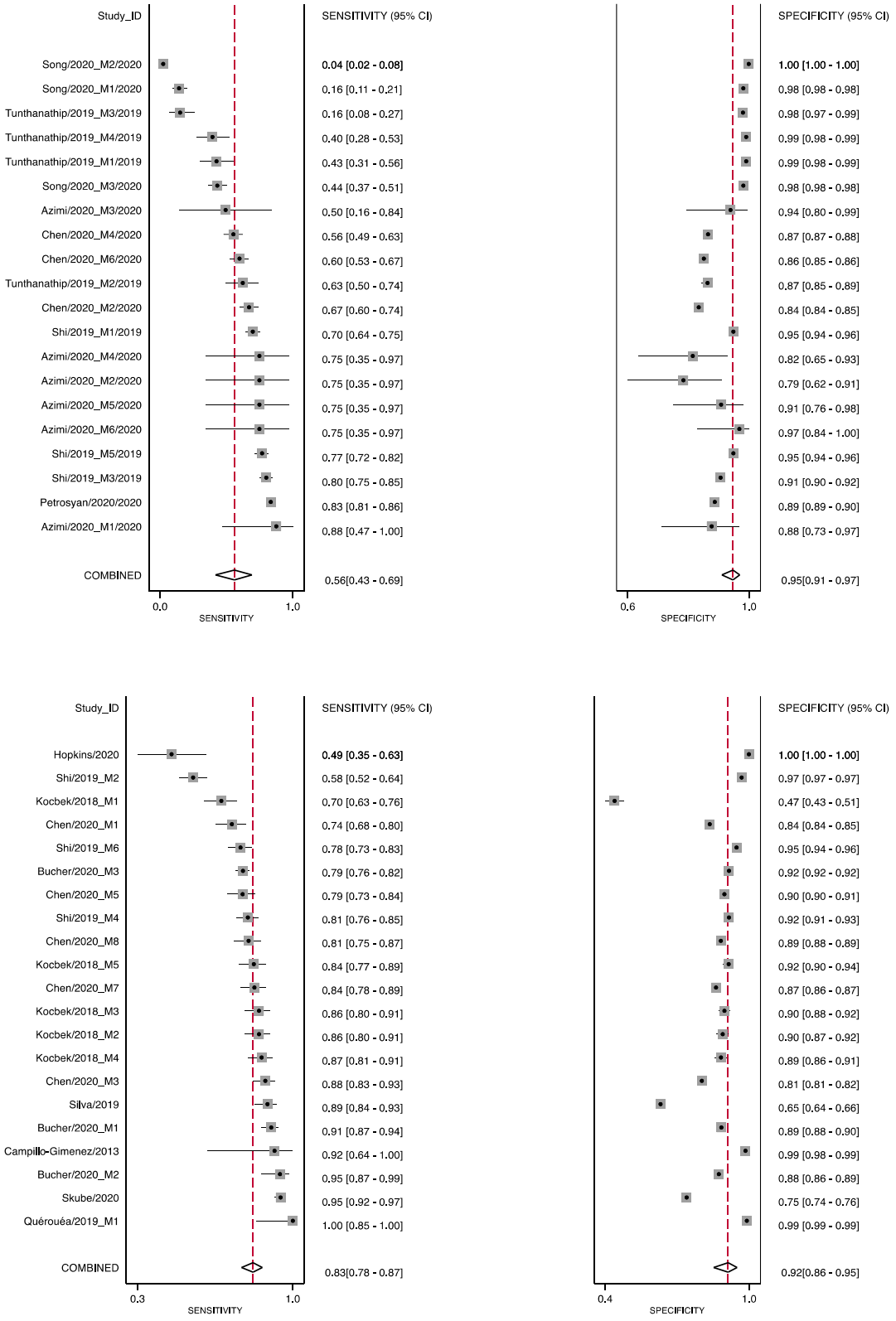

Fig. 2. Forest plot of sensitivity and specificity for structured data-based algorithms (2A) and mixed data-based algorithms (2B).

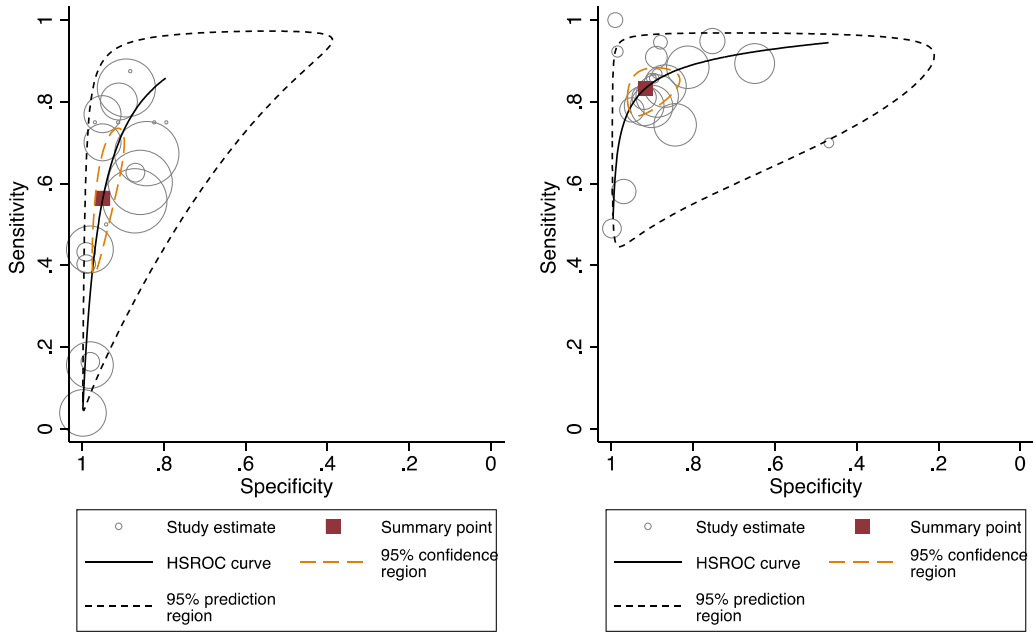

Fig. 3. Hsroc curve for structured data-based algorithms (left) and mixed data-based algorithms (right).

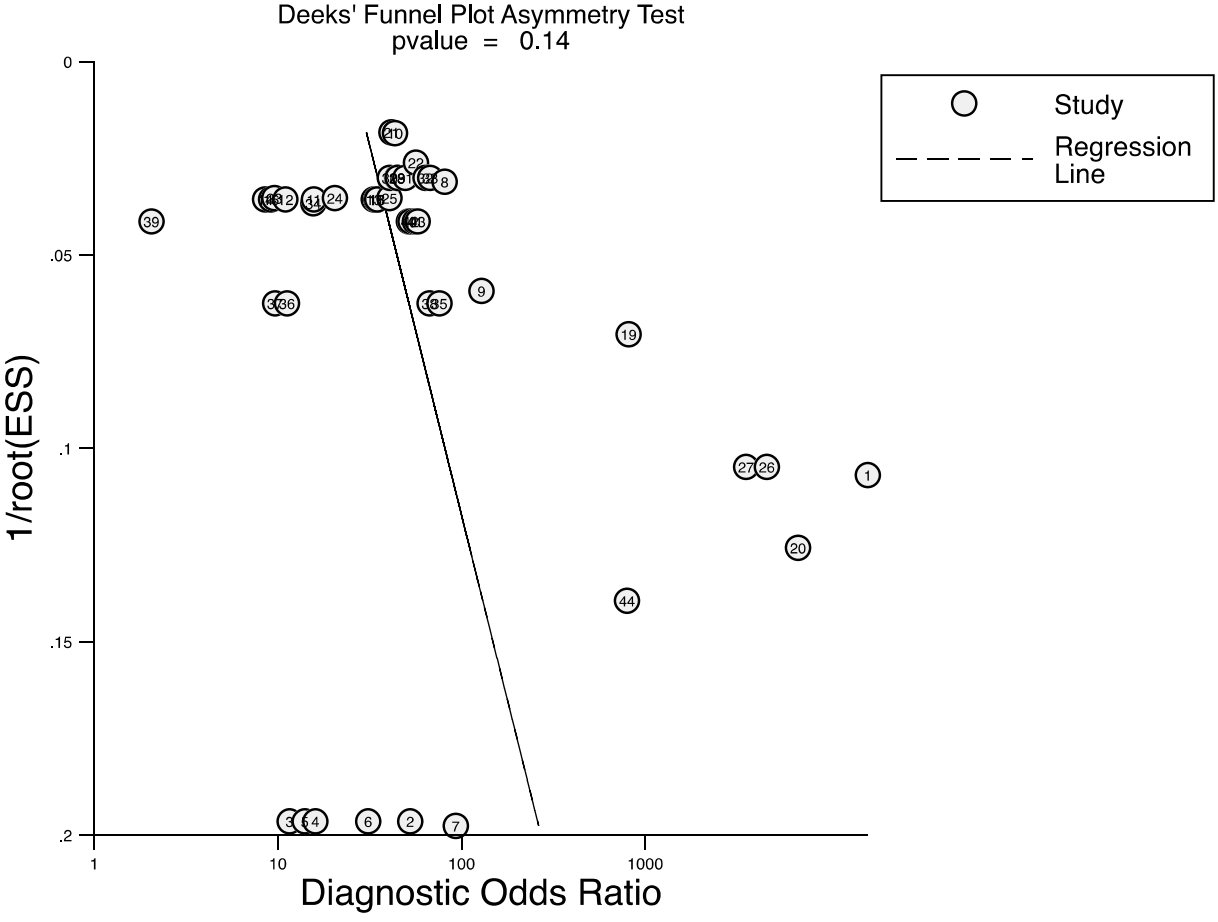

Fig. 4. Deeks funnel plot for publication bias assessment.

Heterogeneity was expected in this review [19,20,54,55]. Sources of heterogeneity were explored with meta-regression and stratified analysis was performed. Other sources, like the threshold effect, could also cause heterogeneity [59]. The threshold in ML is the best score we could choose and set from decision function output to label a sample as a negative or positive class to achieve optimal model performance [60]. In addition, a threshold can be decided for research purposes (high sensitivity/PPV or F1 score) and tuned during model development. In this review, performance estimates were pooled using the HSROC model, considered the most statistically rigorous model to mitigate threshold effect, as recommended by the Cochrane Collaboration Center and the Agency for Healthcare Research and Quality (AHRQ) [18,19,21,59]. Descriptive statistics (median, IQR) of raw data and pooled estimates of HSROC model for algorithm performance measures suggest similar differences in magnitude and direction.

The application of ML technologies into medical research is promising, and validity is still the crucial step for generalizability [61]. Almost all (93.8%) of the algorithms were internally validated, with only a single study providing external validation, using a large cohort from a different healthcare system [10]. The majority (65.6%) of included articles were single-center studies, so the external validation for developed algorithms remains a concern [61]. Clinical implementation of developed algorithms was not explicitly suggested in articles, and knowledge translation studies are still largely needed.

5. Limitations

To the best of our knowledge, this is the first systematic review to summarize the performance of ML algorithms in SSI case detection and prediction. Despite rigorous review steps and applying multiple statistic methodologies, our findings must be carefully interpreted with the following limitations. First, though the conclusion is clear, pooled ML performance estimates are subjected to chances of heterogeneity. We suggest using descriptive statistics of raw data instead. Second, we re-calculated the two-by-two table with reported measures from the included articles, and there might be a chance of misclassification due to rounding. However, we estimate the impact would be minimal compared to the large cardinality of included surgical procedures. Lastly, the stratified analyses of individual ML methodology were built on a limited number of reported studies which may not accurately reflect their general performance.

6. Conclusion

The application of ML algorithms into medical practice has been promising in the past decade. Algorithms developed with mixed-use of structured and textual data provided optimal performance for SSI detection and prediction. However, external validation of developed algorithms is needed for translating current knowledge into clinical practice.

Provenance and peer review

Not commissioned, externally peer-reviewed.

Data sharing statement

Data for this research are available from the corresponding author on reasonable request.

Ethical approval

NA.

Sources of funding

No funding is associated with the collection, analysis and interpretation of data, or with the writing of the manuscript; and in the decision to submit the manuscript for publication.

Author contribution

Author's Contribution: GW conceived this study. GW, CE and DS designed the study. GW retrieved the publications and together with SK and FY conducted the two rounds of review. GW, FY and SK conducted data extraction. GW conducted the data analyses. GW drafted the manuscript and all authors contributed to the revision. All authors agreed on the final version of submission and account for all aspects of this work.

Registration of research studies

This review only included publicly available articles and do not involve human participants.

1. Name of the registry: PROSPERO
2. Unique Identifying number or registration ID: CRD42022339630
3. Hyperlink to your specific registration (must be publicly accessible and will be checked): https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022339630

Guarantor

Guosong Wu.

Consent

NA.

Declaration of competing interest

Dr. Wu is supported by Network of Alberta Health Economists Postdoctoral Fellowship. The remaining authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We thank Dr. Diane L. Lorenzetti for the help of developing search strategy.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.amsu.2022.104956.

References

[1] J. Badia, A. Casey, N. Petrosillo, et al., Impact of surgical site infection on healthcare costs and patient outcomes: a systematic review in six European countries, J. Hosp. Infect. 96 (1) (2017) 1–15.
[2] S.I. Berríos-Torres, C.A. Umscheid, D.W. Bratzler, et al., Centers for disease control and prevention guideline for the prevention of surgical site infection, 2017, JAMA surgery 152 (8) (2017) 784–791.
[3] Canadian Surgical Site Infection Prevention Audit Month – RECAP REPORT, February 2016. https://www.patientsafetyinstitute.ca/en/toolsResources/SSI-Audit-2016/Documents/SSI%20Audit%202016_Recap%20Report%20EN.pdf.
[4] M.S. Morris, R.J. Deierhoi, J.S. Richman, et al., The relationship between timing of surgical complications and hospital readmission, JAMA surgery 149 (4) (2014) 348–354.
[5] W.H. Organization, Global Guidelines for the Prevention of Surgical Site Infection, second ed., 2018, 9789241550475. https://www.who.int/publications/i/item/global-guidelines-for-the-prevention-of-surgical-site-infection-2nd-.

[6] T.G. Weiser, A.B. Haynes, G. Molina, et al., Size and distribution of the global volume of surgery in 2012, Bull. World Health Organ. 94 (3) (2016) 201.

[7] D.A. da Silva, C.S. Ten Caten, R.P. Dos Santos, et al., Predicting the occurrence of surgical site infections using text mining and machine learning, PLoS One 14 (12) (2019), e0226272.

[8] S. Sohn, D.W. Larson, E.B. Habermann, et al., Detection of clinically important colorectal surgical site infection using Bayesian network, J. Surg. Res. 209 (2017) 168–173.

[9] A.M. Darcy, A.K. Louie, L.W. Roberts, Machine learning and the profession of medicine, JAMA 315 (6) (2016) 551–552.

[10] B.T. Bucher, J. Shi, J.P. Ferraro, et al., Portable automated surveillance of surgical site infections using natural language processing: development and validation, Ann. Surg. 272 (4) (2020) 629–636.

[11] M.A. Bartz-Kurycki, C. Green, K.T. Anderson, et al., Enhanced neonatal surgical site infection prediction model utilizing statistically and clinically significant variables in combination with a machine learning algorithm, Am. J. Surg. 216 (4) (2018) 764–777.

[12] H.J. Murff, F. FitzHenry, M.E. Matheny, et al., Automated identification of postoperative complications within an electronic medical record using natural language processing, JAMA 306 (8) (2011) 848–855.

[13] G. Wu, S. Khair, F. Yang, C. Eastwood, The Performance of Machine Learning Algorithm in Surgical Site Infections Case Identification and Prediction, a Systematic Review Protocol, PROSPERO, 2022.

[14] M.D. McInnes, D. Moher, B.D. Thombs, et al., Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement, JAMA 319 (4) (2018) 388–396.

[15] A.J. Viera, J.M. Garrett, Understanding interobserver agreement: the kappa statistic, Fam. Med. 37 (5) (2005) 360–363.

[16] A. Schwartz, Diagnostic test calculator. http://araw.mede.uic.edu/cgi-bin/testcalc.pl?DT=32&Dt=173&dT=172&dt=8897&2x2=Compute.

[17] P.F. Whiting, A.W. Rutjes, M.E. Westwood, et al., QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies, Ann. Intern. Med. 155 (8) (2011) 529–536.

[18] P. Macaskill, C. Gatsonis, J. Deeks, et al., Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy, Version, 2010.

[19] J. Lee, K.W. Kim, S.H. Choi, et al., Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-part II. Statistical methods of meta-analysis, Korean J. Radiol. 16 (6) (2015) 1188–1196.

[20] K.W. Kim, J. Lee, S.H. Choi, et al., Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-part I. General guidance and tips, Korean J. Radiol. 16 (6) (2015) 1175–1187.

[21] T.A. Trikalinos, C.M. Balion, C.I. Coleman, et al., meta-analysis of test performance when there is a "gold standard", J. Gen. Intern. Med. 27 (1) (2012) 56–66.

[22] J.J. Deeks, P. Macaskill, L. Irwig, The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed, J. Clin. Epidemiol. 58 (9) (2005) 882–893.

[23] D. Stengel, K. Bauwens, J. Sehouli, et al., Original Paper: a likelihood ratio approach to meta-analysis of diagnostic studies, J. Med. Screen 10 (1) (2003) 47–51.

[24] A. Van Esbroeck, I. Rubinfeld, B. Hall, et al., Quantifying surgical complexity with machine learning: looking beyond patient factors to improve surgical models, Surgery 156 (5) (2014) 1097–1105.

[25] A.K. Gowd, A. Agarwalla, N.H. Amin, et al., Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty, J. Shoulder Elbow Surg. 28 (12) (2019) e410–e421.

[26] R.W. Grundmeier, R. Xiao, R.K. Ross, et al., Identifying surgical site infections in electronic health data using predictive models, J. Am. Med. Inf. Assoc. 25 (9) (2018) 1160–1166.

[27] Z. Hu, G.J. Simon, E.G. Arsoniadis, et al., Automated detection of postoperative surgical site infections using supervised methods with electronic health record data, Stud. Health Technol. Inf. 216 (2015) 706.

[28] Z. Hu, G.B. Melton, N.D. Moeller, et al., Accelerating Chart Review Using Automated Methods on Electronic Health Record Data for Postoperative Complications, American Medical Informatics Association, 2016, p. 1822.

[29] C. Ke, Y. Jin, H. Evans, et al., Prognostics of surgical site infections using dynamic health data, J. Biomed. Inf. 65 (2017) 22–33.

[30] P. Mandagani, S. Coleman, A. Zahid, et al., Machine Learning Models for Surgical Site Infection Prediction, 2016.

[31] K. Merath, J.M. Hyer, R. Mehta, et al., Use of machine learning for prediction of patient risk of postoperative complications after liver, pancreatic, and colorectal surgery, J. Gastrointest. Surg. 24 (8) (2020) 1843–1851.

[32] J.D. Michelson, J.S. Pariseau, W.C. Paganelli, Assessing surgical site infection risk factors using electronic medical records and text mining, Am. J. Infect. Control 42 (3) (2014) 333–336.

[33] P.C. Sanger, G.H. van Ramshorst, E. Mercan, et al., A prognostic model of surgical site infection using daily clinical wound assessment, J. Am. Coll. Surg. 223 (2) (2016) 259–270, e2.

[34] F. Shen, D.W. Larson, J.M. Naessens, et al., Detection of surgical site infection utilizing automated feature generation in clinical notes, Journal of healthcare informatics research 3 (3) (2019) 267–282.

[35] C. Soguero-Ruiz, W.M. Fei, R. Jenssen, et al., Data-driven Temporal Prediction of Surgical Site Infection, American Medical Informatics Association, 2015, p. 1164.

[36] A.S. Strauman, F.M. Bianchi, K.Ø. Mikalsen, et al., Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks, IEEE (2018) 307–310.

[37] C.P. Thirukumaran, A. Zaman, P.T. Rubery, et al., Natural language processing for the identification of surgical site infections in orthopaedics, J. Bone Jt. Surg. Am. Vol. 101 (24) (2019) 2167.

[38] G.B. Weller, J. Lovely, D.W. Larson, et al., Leveraging electronic health records for predictive modeling of post-surgical complications, Stat. Methods Med. Res. 27 (11) (2018) 3271–3285.

[39] A.B. Chapman, D.L. Mowery, D.S. Swords, et al., Detecting evidence of intra-abdominal surgical site infections from radiology reports using natural language processing, American Medical Informatics Association (2017) 515.

[40] A.V. Karhade, M.E. Bongers, O.Q. Groot, et al., Can natural language processing provide accurate, automated reporting of wound infection requiring reoperation after lumbar discectomy? Spine J. 20 (10) (2020) 1602–1609.

[41] W. Chen, Z. Lu, L. You, et al., Artificial intelligence–based multimodal risk assessment model for surgical site infection (AMRAMS): development and validation study, JMIR medical informatics 8 (6) (2020), e18186.

[42] J. Shi, S. Liu, L.C. Pruitt, et al., Using Natural Language Processing to Improve EHR Structured Data-Based Surgical Site Infection Surveillance, American Medical Informatics Association, 2019, p. 794.

[43] T. Tunthanathip, S. Sae-Heng, T. Oearsakul, et al., Machine learning applications for the prediction of surgical site infection in neurological operations, Neurosurg. Focus 47 (2) (2019) E7.

[44] P. Kocbek, N. Fijacko, C. Soguero-Ruiz, et al., Maximizing interpretability and cost-effectiveness of Surgical Site Infection (SSI) predictive models using feature-specific regularized logistic regression on preoperative temporal data, Comput. Math. Methods Med. 2019 (2019).

[45] M.L. Ciofi Degli Atti, F. Pecoraro, S. Piga, et al., Developing a surgical site infection surveillance system based on hospital unstructured clinical notes and text mining, Surg. Infect. 21 (8) (2020) 716–721.

[46] B.S. Hopkins, A. Mazmudar, C. Driscoll, et al., Using artificial intelligence (AI) to predict postoperative surgical site infection: a retrospective cohort of 4046 posterior spinal fusions, Clin. Neurol. Neurosurg. 192 (2020), 105718.

[47] B. Campillo-Gimenez, N. Garcelon, P. Jarno, et al., Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France, 2013. IOS Press, Medinfo (2013) 572–575.

[48] K. Azimi, M.D. Honaker, S. Chalil Madathil, et al., Post-operative infection prediction and risk factor Analysis in colorectal surgery using data mining techniques: a pilot study, Surg. Infect. 21 (9) (2020) 784–792.

[49] Y. Petrosyan, K. Thavorn, G. Smith, et al., Predicting Postoperative Surgical Site Infection with Administrative Data: a Machine Learning Algorithm, 2020.

[50] M. Quéroué, A. Lashéras-Bauduin, V. Jouhet, et al., Automatic detection of surgical site infections from a clinical data warehouse, arXiv preprint arXiv:190907054 (2019).

[51] J. Song, E. Sanabria-Buenaventura, B. Cohen, et al., Predictive Models for Surgical Site Infection (SSI) in Patients with a Permanent Pacemaker (PPM) Using Machine Learning Methods, Authorea Preprints, 2020.

[52] S.J. Skube, Z. Hu, G.J. Simon, et al., Accelerating surgical site infection abstraction with a semi-automated machine-learning approach, Ann. Surg. (2020).

[53] A.W. Rutjes, J.B. Reitsma, J.P. Vandenbroucke, et al., Case–control and two-gate designs in diagnostic accuracy studies, Clin. Chem. 51 (8) (2005) 1335–1341.

[54] J.G. Lijmer, P.M. Bossuyt, S.H. Heisterkamp, Exploring sources of heterogeneity in systematic reviews of diagnostic tests, Stat. Med. 21 (11) (2002) 1525–1537.

[55] B. Dwamena, R. Sylvester, Carlos R. midas, Meta-analysis of diagnostic accuracy studies, *Available at:) View in Article*, http://fmwww.bc.edu/repec/bocode/m/midas.pdf, 2009. (Accessed 8 February 2017).

[56] R.P. Merkow, M.H. Ju, J.W. Chung, et al., Underlying reasons associated with hospital readmission following surgery in the United States, JAMA 313 (5) (2015) 483–495.

[57] I.J.B. Young, S. Luz, N. Lone, A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis, Int. J. Med. Inf. 132 (2019), 103971.

[58] B. Mahesh, Machine learning algorithms-A review, International Journal of Science and Research (IJSR)[Internet] 9 (2020) 381–386.

[59] P. Cronin, A.M. Kelly, D. Altaee, et al., How to perform a systematic review and meta-analysis of diagnostic imaging studies, Acad. Radiol. 25 (5) (2018) 573–593.

[60] Z. Omary, F. Mtenzi, Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning, International Journal for Infonomics (IJI) 3 (3) (2010) 314–325.

[61] A. Scardoni, F. Balzarini, C. Signorelli, et al., Artificial intelligence-based tools to control healthcare associated infections: a systematic review of the literature, Journal of infection and public health 13 (8) (2020) 1061–1077.