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

Guosong Wu, Shahreen Khair, Fengjuan Yang, Cheligeer Cheligeer, Danielle Southern, Zilong Zhang, Yuanchao Feng, Yuan Xu, Hude Quan, Tyler Williamson, Cathy A. Eastwood

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.
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 I²) 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
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...

Fig. 1. Selection of articles for Review.
Abbreviations: SSI, surgical site infections; ML, machine learning.
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/Detector) mixed-use of structured

---

**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     | Italy    | Pediatric  | Retro        | Single  | Consecutive     | Mean – 60.25 | 106 (51.5)    | Mixed        | 30                | NR             |
| Azimi/2020    | USA      | NR         | Retro        | Single  | Consecutive     | Mean – 25.3  | 71 (43.8)     | Mixed        | 30                | NR             |
| Bucher/2020   | USA      | NR         | Retro        | Multiple| Random          | Mean – 53, SD| 11,077        | Colorectal    | 30                | G              |
| Chen/2020     | China    | Adult      | Retro        | Multiple| Consecutive     | Median – 54.3| 13,293        | Mixed        | 30                | NR             |
| Hopkins/2020  | USA      | NR         | Retro        | Single  | Consecutive     | Median – 57.5| 2104 (52)     | Orthopedic    | 90                | P              |
| Karhade/2020  | USA      | Adult      | Retro        | Multiple| Consecutive     | Median – 47, IQR [37,59] | 2628 (44.8) | Orthopedic    | 90                | P              |
| Merath/2020   | USA      | Adult      | Retro        | Multiple| Consecutive     | Median – 66, IQR (57, 75) | NR    | Mixed        | 30                | G              |
| Petrozyan/2020| Canada   | Adult      | Retro        | Single  | Consecutive     | Mean – 56.7  | 4540 (G)      | Mixed        | 30                | G              |
| Skube/2020    | USA      | NR         | Retro        | Single  | Consecutive     | Mean – 54, SD = 17 | 5094 [45] | Mixed        | 30                | NR             |
| Song/2020     | USA      | Mixed      | Retro        | Multiple| Consecutive     | Mean – 54.7  | 1786 (56.3)   | Cardiac and vascular | 30                | NR             |
| Goved/2019    | USA      | NR         | Retro        | Multiple| Consecutive     | Mean – 69.5, SD = 9.6 | 7493 (43.8) | Mixed        | 30                | NR             |
| Quereuza/2019 | France   | NR         | Retro        | Single  | Consecutive     | Mean – 69.5, SD = 9.6 | 7493 (43.8) | Mixed        | 30                | NR             |
| Shen/2019     | USA      | NR         | Retro        | Single  | Consecutive     | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Shi/2019      | USA      | NR         | Retro        | Multiple| Consecutive     | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Silva/2019    | Brazil   | Pediatric  | Retro        | Single  | Consecutive     | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Thirukumaran/2019 | USA | NR | CC | Single | Consecutive | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Tunthanathip/2019 | Thailand | NR | Retro | Single | Random       | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Grundmeier/2018 | USA | Pediatric | Retro | Multiple | Consecutive    | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Kocbek/2018   | Slovenia | NR         | Retro        | Single  | Consecutive     | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Strauman/2018 | Norway   | NR         | Retro        | Single  | Consecutive     | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Weller/2018   | USA      | NR         | Retro        | Single  | Consecutive     | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Chapman/2017  | USA      | NR         | Retro        | Multiple| Consecutive     | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Sohn/2017     | USA      | NR         | Retro        | Single  | Consecutive     | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Hu/2016       | USA      | Pediatric  | Retro        | Single  | Consecutive     | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Ke/2016       | USA      | NR         | Pro          | Single  | Consecutive     | Mean – 54.3  | 47 (50.8)     | Mixed        | 30                | NR             |
| Mandagani/2016 | USA | NR | CC | NR | NR | NR | 542 (63.1) | NR | NR | Gastrointestinal  |
| Sanger/2016   | USA      | NR         | Pro          | Single  | Random          | Mean – 54.3  | 542 (63.1) | General- abdominal | 30                | NR             |
| Hu/2015       | USA      | NR         | Retro        | Single  | Consecutive     | Mean – 54.3  | 542 (63.1) | General- abdominal | 30                | NR             |
| Soguero-Ruiz/2015 | USA | Norway | CC | Single | Random       | Mean – 54.3  | 542 (63.1) | General- abdominal | 30                | NR             |
| Esbroeck/2014 | USA      | Adult      | Retro        | Multiple| Consecutive     | Mean – 54.3  | 542 (63.1) | General- abdominal | 30                | NR             |
| Michelson/2014 | USA | Adult | Retro | Multiple | Consecutive   | Mean – 54.3  | 542 (63.1) | General- abdominal | 30                | NR             |
| Campillo-Gimenez/2013 | France | Adult | Retro | Single | Consecutive    | Mean – 54.3  | 542 (63.1) | General- abdominal | 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.

---
| Study ID | Data Type | Reference Standard | ML Technique | Model Type | Model Threshold | Model Training | Model Validation |
|----------|-----------|--------------------|--------------|------------|----------------|----------------|------------------|
| Atti/2020 [43] | S-EMR, FT | Hospital surveillance program | RE | D | NR | T (2,944) | T (0.61) | NR | NR | NA |
| Azimi/2020 [46] | S-EMR | NR | BN, DT, SVM, ANNs, RF | P | NR | T (208) | T (8.65) | NR | NR | NA |
| Bucher/2020 [10] | S-EMR | NR | NER | D | NR | 4574 | 55.7 | 17,210 | 793 | 4.61 |
| Chen/2020 [39] | S-EMR, FT | National surveillance data (NNS) | LR, RF, DT, ANNs | P | NR | 17,597 | 1.15 | 4014 | 43 | 1.07 |
| Hopkins/2020 [44] | FT | Chart review | ANNs | P | NR | 3034 | T (1.48) | 1012 | NR | NA |
| Karhade/2020 [38] | FT | Chart review | BC | 0.05, 0.1, 0.5 | 4483 | 46 | 1.03 | 1377 | 16 | 1.16 |
| Merath/2020 [10] | S-EMR | Hospital surveillance program | LR, RF, DT, ANNs | P | NR | 10,046 | 556 | 5.53 | 4305 | 239 | 5.55 |
| Skube/2020 [50] | S-EMR | ACS-NSQIP | LR, DT, SVM | P | NR | 6188 | T (205) | 5132 | 161 | 3.14 |
| Song/2020 [49] | S-EMR, ADMIN | National surveillance data | LR, RF, SVM | P | NR | 4014 | 1.15 | 1855 | NR | NA |
| Gowd/2019 [23] | S-EMR, FT | Hospital surveillance program | LR | P | NR | 13,697 | NR | 3422 | NR | NA |
| Queroueia/2019 [40] | S-EMR, FT | Hospital surveillance program | LR | D | NR | 1263 | 13.62 | 316 | 36 | 11.39 |
| Shen/2019 [32] | FT | Chart review | DT, SVM, RF | D | NR | T (1,178) | T (6.79) | NR | NR | NA |
| Shi/2019 [40] | S-EMR, FT | Chart review | RF, SVM, LR | P | NR | T (5,795) | T (291) | NR | NR | NA |
| Silva/2019 [7] | S-EMR, FT | Hospital surveillance program | RF, LR, SVM, BN, NC, SGD | P/D | NR | 188 | 1.21 | 12,637 | 202 | 1.60 |
| Thirukumaranan/2019 [35] | ADMIN, S-EMR, FT | Hospital surveillance program | LR | P | NR | 1263 | 13.62 | 316 | 36 | 11.39 |
| Tanthanthip/2019 [41] | S-EMR, FT | Hospital surveillance program | LR | P | NR | T (1,471) | T (4.55) | 295 | NR | NA |
| Grundmeier/2018 [34] | ADMIN, S-EMR, FT | Hospital surveillance program | LR | P | NR | 6871 | 209 | 3.04 | 1039 | 25 | 2.41 |
| Kocek/2018 [42] | S-EMR, FT | Hospital surveillance program | LR, BC | P | Range: 0.171-0.245 | | 909 | 228 | 50 | 21.93 |
| Strauman/2018 [34] | S-EMR, FT | ICD-10 and Procedure codes | ANNs | D | NR | T (883) | T (26.27) | NR | 232 | NA |
| Weller/2018 [36] | S-EMR, FT | Hospital surveillance program | LR, RF, SVM, BN, BC | P | NR | 1051 | 9.71 | 232 | 18 | 7.76 |
| Chapman/2017 [17] | S-EMR, FT | Hospital surveillance program | LR | P | NR | 565 | 100 | NR | NA |
| Sohn/2017 [8] | S-EMR, FT | Hospital surveillance program | BN | D | NR | T (751) | T (8.92) | NR | NR | NA |
| Hu/2016/2016 [26, 28] | S-EMR, FT | ACS-NSQIP | LR | P | NR | T (2,407) | T (8.45) | 163 | NR | NA |
| Ke/2016 [27] | S-EMR, FT | ACS-NSQIP | LR | P | NR | 5280 | 6.36 | 3629 | 157 | 4.33 |
| Mandagani/2016 [28] | S-EMR | Hospital surveillance program | LR, DT | P | NR | T (181) | T (19.62) | NR | NR | NA |
| Sanger/2016 [31] | S-EMR | Hospital surveillance program | LR | P | NR | T (851) | T (19.62) | NR | 229 | NA |
| Hu/2015 [25] | S-EMR, FT | ACS-NSQIP | LR | P | NR | 396 | 6.96 | 2262 | 127 | 5.61 |
| Sogueno-Ruiz/2015 [33] | S-EMR, FT | ICD-10 and Procedure codes | SVM | P | NR | T (1,005) | T (10.65) | NR | NR | NA |
| Esbroek/2014 [32] | S-EMR, FT | ACS-NSQIP | LR | P | NR | T (851) | T (19.62) | NR | NR | NA |
| Michelson/2014 [30] | S-EMR, FT | Hospital surveillance program | LR | P | NR | T (2,407) | T (8.45) | 163 | NR | NA |
| Campillo-Gimenez/2013 [45] | S-EMR, FT | Hospital surveillance program | VSM | D | NR | 3785 | 1.11 | 1225 | NR | NA |
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).

Deeks' Funnel Plot Asymmetry Test
p-value = 0.14

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 [19]. 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 recalculated 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.jamsu.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. Berrios-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-.
