Journal of Infection and Public Health xxx (2020) xxx–xxx

Contents lists available at ScienceDirect

Journal of Infection and Public Health

journal homepage: http://www.elsevier.com/locate/jiph

Review

Artificial intelligence-based tools to control healthcare associated infections: A systematic review of the literature

Alessandro Scardoni[a], Federica Balzarini[a], Carlo Signorelli[a], Federico Cabitza[b], Anna Odone[a,c,*]

[a] School of Medicine, Vita-Salute San Raffaele University, Milan, Italy
[b] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy
[c] Clinical Epidemiology and HTA, IRCCS San Raffaele Scientific Institute, Milan, Italy

ARTICLE INFO

Article history:
Received 28 March 2020
Received in revised form 24 May 2020
Accepted 2 June 2020

Keywords:
Artificial intelligence
Machine learning
Healthcare-associated infections (HAI)
Infection control
Systematic review

ABSTRACT

Background: Healthcare-associated infections (HAIs) are the most frequent adverse events in healthcare and a global public health concern. Surveillance is the foundation for effective HAIs prevention and control. Manual surveillance is labor intensive, costly and lacks standardization. Artificial Intelligence (AI) and machine learning (ML) might support the development of HAI surveillance algorithms aimed at understanding HAIs risk factors, improve patient risk stratification, identification of transmission pathways, timely or real-time detection. Scant evidence is available on AI and ML implementation in the field of HAIs and no clear patterns emerges on its impact.

Methods: We conducted a systematic review following the PRISMA guidelines to systematically retrieve, quantitatively pool and critically appraise the available evidence on the development, implementation, performance and impact of ML-based HAIs detection models.

Results: Of 3445 identified citations, 27 studies were included in the review, the majority published in the US (n = 15, 55.6%) and on surgical site infections (SSI, n = 8, 29.6%). Only 1 randomized controlled trial was included. Within included studies, 17 (63%) ML approaches were classified as predictive and 10 (37%) as retrospective. Most of the studies compared ML algorithms' performance with non-ML logistic regression statistical algorithms, 18.5% compared different ML models' performance, 11.1% assessed ML algorithms' performance in comparison with clinical diagnosis scores, 11.1% with standard or automated surveillance models. Overall, there is moderate evidence that ML-based models perform equal or better as compared to non-ML approaches and that they reach relatively high-performance standards. However, heterogeneity amongst the studies is very high and did not dissipate significantly in subgroup analyses, by type of infection or type of outcome.

Discussion: Available evidence mainly focuses on the development and testing of HAIs detection and prediction models, while their adoption and impact for research, healthcare quality improvement, or national surveillance purposes is still far from being explored.

Contents

* Corresponding author at: School of Medicine, University Vita-Salute San Raffaele, Via Olgettina, 58, 20132 Milan, Italy.
E-mail address: odone.anna@hsr.it (A. Odone).

https://doi.org/10.1016/j.jiph.2020.06.006
1876-0341/

Introduction

Healthcare-associated infections (HAIs) – intended as infections occurring during the process of care – are the most frequent adverse events in healthcare, a major threat to patient safety and a global public health concern [1]. The impact of HAIs is reflected in considerable clinical and financial burden in terms of prolonged hospital stay, excess death and long-term disability, increased antimicrobial resistance, increased direct costs for health systems and financial loss for patients and families [2]. It is estimated that more than 2.6 million new cases of healthcare-associated infections occur every year in Europe, with a cumulative burden higher than all other reported infectious diseases [3,4]. The alarming burden of HAIs has recently been highlighted in South East Asia and Africa [5,6]. In the US, 1 in 31 patients per day develop at least one healthcare-associated infection, overall responsible for 72,000 deaths per year [46]. Meta-analyses estimated in almost $10 billion the annual cost in the US for the cumulative burden of: central line-associated bloodstream infections (CLABSI), ventilator-associated pneumonia (VAP), surgical site infections (SSI), *Clostridium difficile* infections (CDI) and catheter-associated urinary tract infections (CAUTI) [8].

Surveillance of HAIs is the foundation for organizing, implementing, and maintaining effective infection prevention and control programs. Objectives of HAIs surveillance are: to quantify rates of infections and compare them within/between healthcare facilities, engage clinical teams to adopt best practices, introduce evidence-based and cost-effective interventions to reduce HAI and to identify priority areas where to allocate resources. Surveillance data is used to quantify and monitor HAIs burden, to detect outbreaks, to identify risk factors, to plan, implement and evaluate control interventions, to identify areas for improvement, and to meet reporting mandates [13]. Various surveillance methods have been recommended and validated [9], including continuous surveillance, active/passive surveillance, prevalence surveys, alert-based surveillance, all of which, with different characteristics and at different rates are labor intensive, costly and time consuming [10].

The advances in Information Technologies (IT) and the progressive digitalization of health data offer new tools and potential for the healthcare sector, including for the automation of HAIs surveillance [11]. As recently outlined, the availability of different sources' electronic health data might boost electronic HAI surveillance systems on at least three different levels: (i) enhancing the reliability, efficiency and standardization of surveillance practices [11], (ii) reducing costs and saving times, and (iii) allowing real-time analysis and action [12].

Although automated and semiautomated HAI surveillance systems are traditionally based on fix and a priori defined classification algorithms or simple rule-based decision trees, new evidence suggest that, Artificial Intelligence (AI) and machine learning – the latter intended as an umbrella term for a wide and heterogeneous set of statistical and computational techniques adopted and applied to build AI systems (please refer to Box 1 for technical explanations of artificial intelligence and machine learning models) can support the development of HAI surveillance algorithms [11]. In broad terms, Machine Learning (ML) refer to the iterative and automatic optimization of mathematical models that fits the available data with progressive accuracy. Building on the theoretical concepts outlined in Box 1, its application to infection prevention and control can lead to an improved understanding of HAIs risk factors, improved patient risk stratification, identification of transmission pathways, as well as timely or real-time detection and control. Despites such promising approach, scant evidence is available on the literature on ML implementation in the field of HAIs and no clear patterns emerges on its impact.

Aim of the current study is to collect and summarize the available evidence on the application and impact of Artificial Intelligence to HAIs control. Specific objectives are: to systematically retrieve (i) experiences of AI-based HAIs detection, (ii) their performance measures, as compared to traditional manual or automated detection methods, (iii) to pool and critically appraise the available evidence on the topic, outlining potential strengths and pitfalls and highlighting current gaps in knowledge.

Methods

As done before [23], the review's methods were defined in advance following the Prepared Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [24].

Criteria for considering studies

We included publications that reported on the use of machine learning-based tools to detect and control HAIs. All healthcare-associated infections were considered and no restrictions were applied by type of healthcare facility. Only studies reporting original data were included. Eligible study designs included clinical trials, prospective cohort, retrospective cohort and case-control studies. Literature reviews were screened to retrieve relevant primary data. Inclusion was restricted to full text papers; conference abstracts, posters and study protocols were excluded. Outcomes of interest included all possible performance measures, as well as all possible clinical, organizational and economic outcomes.

Search methods for identification of studies

Studies were identified by searching the electronic databases Medline and Embase. The search strategy was first developed in Medline using a combination of free text and Mesh terms, and then adapted for use in the other databases. Complete search strategies are available in Appendix A. Further studies were retrieved from manual reference listing of relevant articles and consultation with experts in the field. Studies published in English through June 2018 were included.

Box 1: Artificial intelligence and machine learning models

Artificial Intelligence (AI) is a term of great rethorical (and evocative) power [13], but a very low descriptive one, despite its wide use. In this paper we will refer to the so called "narrow AI", which regards computational systems developed to execute specific and circumscribed tasks as much as (or even more) effectively than human performers [21], and definitely more efficiently than humans. Many recent AI systems are built by means of Machine Learning (ML). This latter is an umbrella term for a wide and heterogeneous set of statistical and computational techniques that are usually applied to build (narrow) "AI systems" that exhibit very good performance in tasks involving pattern matching and signal recognition, including image recognition. Despite their great diversity, the element that is common to all ML methods is the iterative and automatic optimization of a mathematical model that fits (i.e., explains, reproduces, interpolates) the available data, the so called training dataset: for this reason, ML is an approach that is based on the available data rather than on explicit and formal representations of either declarative and procedural knowledge (rules and algorithms). The tasks where ML models achieve high performance can be divided in either ***discriminative*** tasks, that is regarding the classification of a new instance of data on the basis of similar data in the training set; or ***regressive*** tasks, when the model is aimed at estimating an unknown numerical value of a data instance. In medical terms, discriminative models can be mainly used to support diagnostic reasoning; regression models can be useful for prognostic and therapeutic purposes. Discriminative model can be further divided according to whether they work on data that have been previously labeled by domain experts or not: in the former case, the models are said to be **supervised**; in the latter case they are **unsupervised** (like in case of clustering algorithms).

Despite the existence of many models, a small number of them usually result to outperform the others: to identify these models, a recent survey has tested an impressive number of different classifiers ($n = 179$) on a likewise impressive number of different data sets ($n = 121$) and concluded that random forests and support vector machines (with Gaussian kernel) are usually the best performing models [16]. Although accuracy (and hence the complementary concept, error rate) is an important feature of ML models, these models are usually developed by finding an acceptable compromise between accuracy and complexity, as depicted in Fig. 1, that is between the problem of underfitting, which occurs when the model is too simplistic to get the intrinsic complexity of both the training data and any possible new data and is characterized by high "bias"; and overfitting, that is the opposite condition when the model "mirrors" the training data too closely but generalizes poorly on new (unseen) data, that is when bias is relatively low but variance too high to yield value in real-world settings [17–20]. Recently, the medical community has also emphasized the importance to consider other dimensions besides accuracy and complexity in the development of ML models, like explainability and causability: the former is the capacity of the AI system to provide credible explanations of the advice given so as to "open" the black-box of models that would be inscrutable otherwise; the latter regards the quality of these explanations to allow "human expert achieve a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use" [22].

Data collection and analysis

Identified studies were independently reviewed for eligibility by two authors (AS, FB) in a two-step based process; a first screening was performed based on title and abstract while full texts were retrieved for the second screening. At both stages disagreements by reviewers were resolved by consensus and consultation with senior authors (AO, FC). Data were extracted by two authors (AS, FB) supervised by a third and fourth author (AO, FC) using a standardized data extraction spreadsheet. The data extraction spreadsheet was piloted on 10 randomly selected papers and modified accordingly. Data extraction included: authors' affiliation, journal, publication year, country of studies' implementation, study design, study setting, study period, type of infection, sample size, machine learning model (intervention), comparison model, information on analysis performed, outcomes of interest, prediction metrics and results.

Analysis and quality appraisal

We performed descriptive analysis to report the characteristics of included studies. Variables' categories regrouping was carried out as following: authors' affiliation was categorized into clinical departments and/or information technology (IT) departments; information on private sector involvement in the authorship was acknowledged. Study setting was divided into surgery and emergency departments, intensive care units (ICU) or general hospital impatient setting. ML models were categorized in *predictive* – helpful to detect real-time patients' risk of HAI – and *retrospective* – helpful for surveillance and epidemiological analysis.

With regard to the pre-specified outcomes of interest we quantitatively retrieved:

- HAIs surveillance models' performance measures, expressed as: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under the receiver operating characteristic curve (AUROC), accuracy, precision, and other performance measures;
- HAIs' clinical and organizational control indicators, expressed as: reduced HAIs' incidence, prevalence, morbidity, mortality, impatient length of stay and costs.

Both performance measures and control indicators were pooled using ranges and average values, grouped by type of infection (HAI in general, CLABSI, SSI, CDI, CAUTI, VAP and other specific infections) and compared differentiating between ML and non-ML-based surveillance models. For studies that presented a comparison of the performance of different ML models, we a priori decided to select and extract data referring to the best performing algorithm.

Anticipating variability between studies, and depending on data availability, we planned to apply – where relevant and possible – random effects analyses to acquire pooled performance and effectiveness estimates for ML-based vs. non-ML based surveillance systems [25]. Included studies quality appraisal was carried out applying: the Newcastle-Ottawa Scale (NOS) [26] for non-randomized studies and the Cochrane Collaboration's tool for randomized studies [27]. Included studies' quality was not set as exclusion criteria. Disagreements by reviewers were resolved by consensus.

Results

We identified 3445 citations by searching the selected databases and listing references of relevant articles. After removing duplicates, 2873 records were retrieved. Papers were screened and selected as illustrated in Fig. 2, resulting in 27 studies meeting our a priori defined inclusion criteria and ultimately included in the review.

The 27 included studies corresponded to 26 different study populations, as two papers referred to the same study [28,29]. Characteristics of included studies are reported in Table 1. Included studies were carried out in 9 countries, the majority in the

Table 1
Characteristics of the included studies.

Ref & year	Country	Affiliation	Study setting	Study population	Sample size	Study period	Study design	Objective	Infection type	Analysis	Comparison	Outcomes
Beeler 2018 [30]	USA	Clinical D. & IT D. & Private Sector	Hospital	Neonatal and pediatric patients	70,218	January 1st, 2013- May 31st, 2016	Retrospective cohort	Evaluate and validate a machine learning as an accurate model to predict the risk of CLABSI in real time	CLABSI	Predictive	RF vs LR	AUROC
Branch-Elliman 2015 [31]	USA	Clinical D.	ICU	ICU/ACU patients	43,609 patient-days	March 1st 2013- November 30th 2013	Prospective cohort	To assess the utility of NLP algorithm for identifying indwelling urinary catheter days and CAUTI in a clinical setting	CAUTI	Predictive	NLP-augmented algorithm vs standard surveillance method	Sensitivity SpecificityPPVN PV
Campillo-Gimenez 2013 [45]	France	Clinical D. & IT D.	Surgery	>18y neurosurgery patients	5010	2008–2010	Retrospective cohort	Automated detection strategy for SSI in neurosurgery, based on textual analysis of medical reports stored in a clinical data warehouse	SSI	Retrospective	NLP vs DRG database vs Conventional surveillance	Recall Precision *F*-measureO verload Index
Chang 2011 [46]	Taiwan	Clinical D. & IT D.	Hospital	All impatients	806 HAI 69,032 non HAI (control group)	2004–2005	Retrospective cohort	Development of a scoring system to predict HAI, derived from Logistic Regression and validated by Artificial Neural Network simultaneously	HAI	Predictive	ANN vs LR vs scoring system	Sensitivity SpecificityAccuracy AUROC
Chen 2014 [48]	China	Clinical D.	Hospital	Lung cancer patients	609	January 2005–January 2014	Retrospective cohort	Development of an ANN model to predict nosocomial infection in lung cancer	HAI	Predictive	ANN vs LR	Sensitivity Specificity PPVNPVAUROCLR+ LR−
Cohen 2004 [28]	Switzerland	Clinical D.	Hospital	>48 h hospitalization impatients	683	2002	Retrospective cohort	Apply data mining techniques to detect nosocomial infections	HAI	Retrospective	No	Sensitivity SpecificityA ccuracy
Cohen 2006 [44]	Switzerland	Clinical D. & IT D.	Hospital	Impatients	688	2002	Retrospective cohort	Identification of patients with a high risk of acquiring any kind of nosocomial infection measuring the performance of a support vector algorithm	HAI	Retrospective	ML vs ML (5 model)	Sensitivity SpecificityAccuracy CWA
Cohen 2008 [29]	Switzerland	Clinical D.	Hospital	>48 h hospitalization impatients	683 cases and 49 variables	2002	Retrospective cohort	Apply data mining techniques to detect nosocomial infections	HAI	Retrospective	No	Sensitivity SpecificityA ccuracy
Desautels 2016 [32]	USA	Clinical D. & IT D. & Private Sector	ICU	>15 y ICU patients	22,853 ICU stays	2001–2012	Retrospective cohort	Compare the machine learning sepsis prediction with existing sepsis scoring systems	Sepsis	Predictive	InSight performance vs sepsis scoring system	Sensitivity SpecificityAccuracy AUROCLR+LR− *F*-measureD iagnostic Odds Ratio

Table 1 (*Continued*)

Ehrentraut 2018 [52]	SwedenFinland	IT D.	Hospital	All impatients	120 patients	Spring 2012	Retrospective cohort	application of support vector machines and gradient tree boosting to detect patient records that include hospital-acquired infections	HAI	Retrospective	ML vs ML	Precision Recall *F*-measure
Escobar 2017 [33]	USA	Clinical D.	Hospital	>18y impatients	11,251	2007–2014	Retrospective cohort	Development and validation of CDI predictive models in a large and representative sample of adults	*Clostridium difficile*	Predictive	Automated model vs Basic model	Sensitivity Specificity PPVN-PVAUROCBrierNNE NRI
Gerbier 2011 [7]	France	Clinical D.	ED	Adult patients	100 medical records	January 1, 2008–March 31, 2010	Retrospective cohort	Description and evaluation of a natural language processing system to extract and encode information found in the narrative reports of computerized ED medical records	HAI	Predictive	No	Recall Precision
Gomez-Vallejo 2016 [51]	Spain	Clinical D. & IT D.	Hospital	All impatients	Training set: 2569 samples, 1800 patients test set: 2816 cases	Training set from 01 March 2012 to 23 January 2013 Test set from 30 September 2013 to 31 August 2014	Retrospective cohort	Development of real-time decision support system for automated surveillance of nosocomial infections	HAI	Predictive	No	Accuracy Kappa Value
Haas 2005 [34]	USA	Clinical D. & IT D.	ICU	Neonates	1692 (NICU 1) 1240 (NICU 2)	From march 1, 2001 through January 31, 2003	Retrospective cohort	Development of an automated monitoring system based on a natural language processor to screen for pneumonia in neonates	Nosocomial pneumonia	Retrospective	No	Sensitivity SpecificityPPV NPV
Hu 2015 [35]	USA	Clinical D. & IT D.	Surgery	Surgical patients	6258 procedures (405 SSIs)	April 2011–December 2013	Retrospective cohort	Automated surgical adverse events detection tool and development of machine learning models to retrospectively detect Surgical Site Infections (SSI), to accelerate the process of extracting postoperative outcomes from medical charts	SSI	Retrospective	No	Specificity NPV AUROC
Hu 2016 [36]	USA	Clinical D. & IT D.	Surgery	Surgical patients	Training set 5280 Test set 3629	2011–2014	Retrospective cohort	Development of an automated postoperative complications detection application by using structured electronic health record (EHR) data	SSI, pneumonia, UTI, sepsis, and septic shock	Retrospective	ML vs ML	N/A

Please cite this article in press as: Scardoni A, et al. Artificial intelligence-based tools to control healthcare associated infections: A systematic review of the literature. J Infect Public Health (2020), https://doi.org/10.1016/j.jiph.2020.06.006

Table 1 (*Continued*)

Ke 2017 [37]	USA	Clinical D. & IT D.	Surgery	Abdominal surgery patients	860	NA	Prospective cohort	Prediction of the onset of SSI using the spatial-temporal matrix data	SSI	Predictive	Support vector regression model vs learning system vs classic regression model	N/A
Kuo 2018 [47]	Taiwan	Clinical D.	Surgery	Head & neck surgery patients	1838	March 2008–February 2017	Retrospective cohort	Comparison of ANN and logistic regression model to predict SSI	SSI	Predictive	ANN vs LR	Sensitivity Specificity Accuracy AUROCBrier DXY
Oh 2018 [38]	USA	Clinical D. & IT D. & Private Sector	Hospital	Adult impatients	191,014 UM 65,718 MGH	UM: January 1, 2010–January 1 2016MGH: June 1 2012–June 1 2014	Retrospective cohort	Evaluate the application to different patient populations of a generalizable machine-learning approach to using the structured data in an EHR to build a CDI risk stratification model tailored to an individual facility	*Clostridium difficile*	Predictive	No	Sensitivity SpecificityPPV AUROC
Parreco 2018 [39]	USA	Clinical D.	ICU	All impatients	57,786	2001–2012	Retrospective cohort	Comparison of machine learning techniques for predicting central line associated bloodstream infection (CLABSI)	CLABSI	Predictive	ML vs ML	Sensitivity Specificity PPVNPVAccuracy AUROCP recision
Sanger 2016 [53]	USA & Netherlands	Clinical D. & IT D. & Private Sector	Surgery	Abdominal surgery patients	851	NA	Prospective cohort	Employ machine learning techniques to develop and test SSI classifiers	SSI	Predictive	No	Sensitivity SpecificityPPV NPV
Savin 2018 [53]	Russia	Clinical D. & IT D.	ICU	ICU >48 h stay	2324	October 1 2010–June 30 2017	Prospective cohort	Identify healthcare-associated ventriculitis and meningitis risk factors using tree based machine learning algorithms	Healthcare-associated ventriculitis and meningitis	Predictive	XgBoost vs LR	PPV NPVAUROCRecall Precision *F*-measure

Table 1 (*Continued*)

Shimabukuro 2017 [40]	USA	Clinical D. & IT D. & Private Sector	ICU	Medical-surgical ICU patients	67 intervention 75 control	Dec 2016–Feb 2017	RCT	Prediction of sepsis	Sepsis	Predictive	ML vs sepsis scores	Hospital LOS ICU LOSI n-hospital mortality rate Sensitivity SpecificityA UROC
Soguero-Ruiz 2015 [49]	Norway	Clinical D. & IT D.	Surgery	Surgical patients	101 cases and 904 controls	NA	Retrospective cohort	Development of a model for real time prediction and identification of patients at risk for developing SSI	SSI	Predictive	ML vs ML	Accuracy
Sohn 2017 [41]	USA	Clinical D. & IT D.	Surgery	Colorectal surgery patients	751 cases	From 2010 to 2012	Retrospective cohort	Assessment of the performance of Bayesian network in abstracting SSI and further evaluation of the potential to identify SSIs from electronic medical records.	SSI	Retrospective	NLP vs LR	AUROC
Taylor 2018 [42]	USA	Clinical D.	ED	Adult impatients	55,365	March 2013–May2016	Retrospective cohort	Selection of best performing ML algorithm	UTI	Retrospective	ML vs UTI diagnosis	Sensitivity SpecificityA ccuracy
Weller 2017 [43]	USA	Clinical D. & IT D. & Private Sector	Surgery	Colorectal surgery patients	4773	2010–2014	Retrospective cohort	Prediction and detection of occurrence of complications of colorectal surgery	SSI	Predictive	ML vs LR	AUROC

(ED) Emergency Department, (ICU) Intensive Care Unit, (ACU) Acute Care Unit, (Clinical D.) Clinical Department, (IT D.) Information Technology Department, (LOS) Length of Stay, (CLABSI) Central Line-associated Bloodstream Infection, (CAUTI) Catheter-associated Urinary Tract Infections, (UTI) Urinary Tract Infections, (SSI) Surgical Site Infections, (HAI) Healthcare-associated Infections, (RF) Random Forest, (LR) Logistic Regression, (NLP) Natural Language Processing, (ANN) artificial neural network, (SVM) Support-vector machine, (PPV) positive predictive value, (NPV) Negative Predictive Value, (CWA) Mean class-weighted accuracy, (AUROC) Area under the Receiver Operating Characteristic, (LR+) positive likelihood ratio, (LR−) negative likelihood ratio, (NNE) Number of incident cases one would need to evaluate to detect one recurrence, (NRI) Net Reclassification Improvement, (APR) Area under the Precision-Recall Curve.

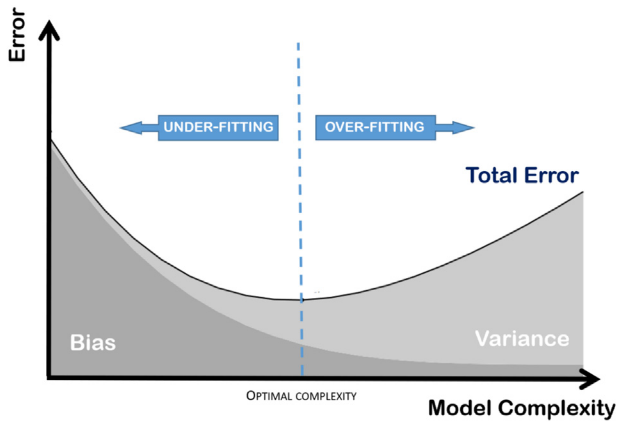

Fig. 1. Optimization between accuracy and complexity of machine learning models.

US ($n = 15$, 55.6%) [30–43], three studies were conducted in Switzerland (11.1%) [28,29,44], 2 in France (7.4%) [7,45], 2 in Taiwan (7.4%) [46,47], and one, respectively, in China [48], Norway [49], Russia [50], Spain [51], Sweden with Finland [52] and The Netherlands with the US [53]. Overall, only 5 studies (18.5%) were conducted in EU countries [7,45,51–53]. Studies were published between 2004 and 2018, with more than one third published from 2017 onwards and one fourth of all included studies ($n = 7$, 25.9%) published in 2018. Most of the studies have at least one author affiliated with a clinical department (96.3%). Nine papers (33.3%) have only authors with clinical department affiliations. Seventy-four per cent of papers resulted from multidisciplinary collaborations ($n = 20$) between clinical and IT researchers. In one fourth of papers ($n = 6$, 22.2%) private companies contributed to the work and were acknowledged in the authorship. Among included studies: 33.3% ($n = 9$) were conducted in surgery departments, 22.2% ($n = 6$) in intensive care units (ICU), 7.4% ($n = 2$) in emergency departments (ED), while 37% ($n = 10$) in general inpatient hospital setting. Two thirds of studies focused on selected types of infections, including: SSI ($n = 8$, 29.6%), healthcare-associated sepsis ($n = 2$, 7.4%), CLABSI ($n = 2$, 7.4%), CDI ($n = 2$, 7.4%), UTI ($n = 1$, 3.7%), CAUTI ($n = 1$, 3.7%), nosocomial pneumonia ($n = 1$, 3.7%), healthcare-associated ventriculitis and meningitis ($n = 1$, 3.7%), while the remaining third ($n = 9$) focused on HAI in general.

The vast majority of studies were retrospective cohorts ($n = 22$, 81.5%), 4 were prospective cohorts [31,37,50,53] and one randomized controlled trial [40]. Units of analysis included: number of patients, number of medical records, patient-days, hospital stay (days), number of procedures. Included studies' sample sizes, differentiating between intervention and control, training test and test sets are reported in Table 1. Sample sizes of studies having patients as unit of analysis ranged from 120 to 256,732 (median 2081).

ML algorithms assessed in included studies varied widely (Table 1): 17 (63%) ML approaches were classified as predictive and 10 (37%) as retrospective. Studies' comparison varied as following: 3 (11.1%) studies assessed ML algorithms' performance in comparison with clinical diagnosis scores [32,40,46], 3 (11.1%) with standard or automated surveillance models [31,33,45], 2 (7.4%) with Diagnosis Related Group (DRG) code detection-based models [42,45]. Most of the studies ($n = 8$, 29.6%) compared ML algorithms' performance with non-AI Logistic Regression statistical algorithms [30,37,41,43,46–48,50]. Five (18.5%) studies compared different ML models' performance [36,39,44,49,52], the remaining studies not providing comparisons (Table 1).

Assessed performance measures were predominantly: specificity (in 16 studies, 59.2%), sensitivity (in15 studies, 55.6%), the area under the receiver operating characteristic curve (AUROC, in 13 studies, 48.1%), accuracy (in 10 studies, 37%), negative predictive value and positive predictive value (in 8 studies, 29.6%), precision ($n = 5$, 18.5%), recall and *F*-measure ($n = 4$, 14.8%). Others considered performance measures are reported in Table 1. Some papers reported on subgroup analysis; 7 papers (25.9%) [35,41,43,46,47,49,53] evaluated the performance by applying different cut-offs or by evaluating the presence of preoperative and postoperative HAIs.

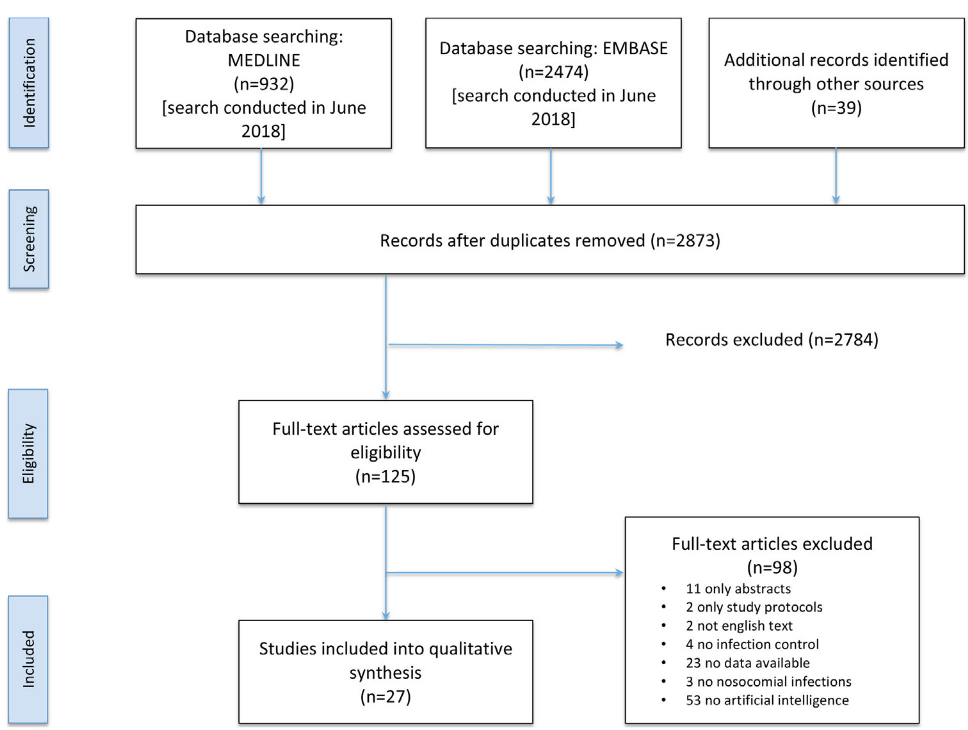

Fig. 2. Screening PRISMA of systematic review.

Table 2
ML-based models for Central Line-associated Bloodstream Infections (CLABSI), sepsis and *Clostridium difficile* infection (CDI) surveillance: performance results.

Ref		Sensitivity	Specificity	PPV	NPV	Accuracy	AUROC	Precision	Sensitivity	Specificity	PPV	NPV	Accuracy	AUROC	Precision
Beeler 2018	CLABSI	ML: random forest					0.87								
Parreco 2018	CLABSI	ML: deep learning 4.0	98.7	4.3	98.6	0.973	0.642	4.3	ML: logistic regression 0.0	100.0	0.0	98.6	0.986	0.722	0.0

Ref		Sensitivity	Specificity	PPV	NPV	Accuracy	AUROC	Precision	AUROC
Beeler 2018	CLABSI								Control: logistic regression 0.79
Parreco 2018	CLABSI	ML: gradient boosted trees 5.3	98.9	7.1	98.6	0.976	0.710	7.1	

Ref		Sensitivity	Specificity	Accuracy	AUROC	LR+	LR−	*F*-measure	APR	Diagnostic odds ratio	Sensitivity	Specificity	Accuracy	AUC/AUROC	LR+	LR−	*F*-measure	APR	Diagnostic odds ratio
Shimabukuro 2017	Sepsis	ML: ML algorithm 90.0	90.0		0.952														
Desautels 2016	Sepsis	ML: InSight (0 h) 80	80	0.80	0.88	3.90	0.25	0.47	0.60	15.51	ML: InSight (4 h) 80	54	0.57	0.74	1.75	0.37	0.30	0.28	4.75

Ref		Sensitivity	Specificity	Accuracy	AUROC	LR+	LR−	*F*-measure	APR	Diagnostic odds ratio
Shimabukuro 2017	Sepsis	Control: Clinical criteria SIRS 59.0MEWS 36.5SOFA 91.0qSOFA 28.8	SIRS 76.4MEWS 66.7SOFA 18.1qSOFA 75.0		SIRS 0.681MEWS 0.524SOFA 0.756qSOFA 0.518					
Desautels 2016	Sepsis	CONTROL: Clinical criteria SIRS 72qSOFA 56MEWS 70SAPS II 75SOFA 80	SIRS 44qSOFA 84MEWS 77SAPS II 52SOFA 48	SIRS 0.47qSOFA 0.80MEWS 0.76SAPS II 0.55SOFA 0.52	SIRS 0.61qSOFA 0.77MEWS 0.80SAPS II 0.70SOFA 0.73	SIRS 1.30qSOFA 3.37MEWS 3.05SAPS II 1.57SOFA 1.55	SIRS 0.63qSOFA 0.53MEWS 0.39SAPS II 0.48SOFA 0.42	SIRS 0.24qSOFA 0.39MEWS 0.40SAPS II 0.27SOFA 0.27	SIRS 0.16qSOFA 0.28MEWS 0.33SAPS II 0.23SOFA 0.28	SIRS 2.06qSOFA 6.33MEWS 7.85SAPS II 3.26SOFA 3.71

Ref		Sensitivity	Specificity	PPV	NPV	AUROC	BRIER	NNE	NRI	Sensitivity	Specificity	PPV	NPV	AUROC	BRIER	NNE	NRI
Escobar 2017	CDI	ML: automated model 79.17	32.04	11.09	93.49	0.605	0.0942	9.02	0.0199	Control: basic model 75.69	41.19	12.11	94.96	0.591	0.0937	8.26	0.0766
Oh 2018	CDI	ML: machine learning algorithm UM 28 MGH 23	UM 95 MGH 95	UM 6 MGH 4		UM 0.82 MGH 0.75											

(ML) Machine learning, (UM) University of Michigan Hospitals, (MGH) Massachusetts General Hospital, (PPV) positive predictive value, (NPV) Negative Predictive Value, (AUROC) Area under the Receiver Operating Characteristic, (LR+) positive likelihood ratio, (LR−) negative likelihood ratio, (F1) *F*-measure, (NNE) Number of incident cases one would need to evaluate to detect one recurrence, (NRI) Net Reclassification Improvement, (APR) Area under the Precision-Recall Curve.

Table 3
ML-based models for Surgical Site Infections (SSI) surveillance: performance results.

Ref		Sensitivity	Specificity	PPV	NPV	Accuracy	AUROC	Recall	Precision	F1	Overload index	DXY	Brier
		ML: Nomindex NLP											
Campillo-Gimenez 2013	SSI							92.3	40.0	55.8	1.6		
		ML: automated supervised learning											
Hu 2015	SSI		93.5 88.8 78.7		98.0 98.5 99.0		0.896						
		ML: ANN											
Kuo 2018	SSI	(A) 61.4 (B) 67.0	(A) 89.0 (B) 95.2			(A) 0.778 (B) 0.757	(A) 0.808 (B) 0.892					(A) 0.615 (B) 0.781	(A) 0.141 (B) 0.090
		ML: NLP Bayesian network											
Sohn 2017	SSI						(1) 0.643 (2) 0.721 (3) 0.799 (4) 0.827						
		ML: LOCF (all tests)											
Soguero-Ruiz 2015	SSI					(A) 0.81 (B) 0.89							
		ML: Naïve Bayes serial features SF classifier (full)											
Sanger 2016	SSI	(C) 42 (D) 69 (E) 80	(C) 91 (D) 78 (E) 64	(C) 53 (D) 43 (E) 35	(C) 87 (D) 91 (E) 93								
		ML: RF											
Weller 2017	SSI						(A) 0.436 (F) 0.465 (G) 0.496 (H) 0.548						

Table 3 (*Continued*)

Ref		Sensitivity	Specificity	PPV	NPV	Accuracy	AUROC	AUROC	AUROC
Campillo-Gimenez 2013	SSI								
Hu 2015	SSI								
Kuo 2018	SSI								
Sohn 2017	SSI								
		ML: warped-GP							
Soguero-Ruiz 2015	SSI					(A) 0.88 (B) 0.90			
		ML: Naïve Bayes serial features SF classifier (simplified)							
Sanger 2016	SSI	(C) 42 (D) 66 (E) 75	(C) 91 (D) 78 (E) 64	(C) 53 (D) 42 (E) 33	(C) 87 (D) 91 (E) 92				
		ML: SVM						ML: AdaBoost	ML: Nbayes
Weller 2017	SSI						(A) 0.553 (F) 0.511 (G) 0.474 (H) 0.494	(A) 0.437 (F) 0.470 (G) 0.511 (H) 0.506	(A) 0.475 (F) 0.450 (G) 0.453 (H) 0.522

Ref		Sensitivity	Specificity	Accuracy	AUROC	Recall	Precision	F1	DXY	Brier	Recall	Precision	F1	Overload index
		Control: conventional surveillance									Control: DRG database			
Campillo-Gimenez 2013	SSI					23.1	100	37.5			84.6	4.8	9.1	21.4
Hu 2015	SSI													
		Control: LR												
Kuo 2018	SSI	(A) 14.4 (B) 22.1	(A) 95.4 (B) 93.3	(A) 0.723 (B) 0.727	(A) 0.694 (B) 0.717				(A) 0.388 (B) 0.433	(A) 0.185 (B) 0.179				
		Control: LR												
Sohn 2017	SSI				0.719									
Soguero-Ruiz 2015	SSI													
Sanger 2016	SSI													
		Control: LassoLR												
Weller 2017	SSI				(A) 0.489 (F) 0.551 (G) 0.563 (H) 0.564									

(A) Pre-operative, (B) post-operative, (C) Higher specificity cutoff, (D) balanced cutoff, (E) higher sensitivity cutoff, (F) postoperative day 0, (G) postoperative day 1, (H) postoperative day 2, (PPV) positive predictive value, (NPV) negative predictive value, (AUROC) Area under the Receiver Operating Characteristic, (F1) *F*-measure.

Please cite this article in press as: Scardoni A, et al. Artificial intelligence-based tools to control healthcare associated infections: A systematic review of the literature. J Infect Public Health (2020), https://doi.org/10.1016/j.jiph.2020.06.006

Table 4
ML-based models for Healthcare Associated Infections (HAIs) and other single infections surveillance: performance results.

Ref	Type of infection	Sensitivity	Specificity	PPV	NPV	Accuracy	CWA	AUROC	LR+	LR−	Recall	Precision	F1	Kappa
		ML: ANN												
Chen 2014	HAI	56.0	85.0	75.7	69.9			0.860	3.73	0.52				
		ML: ANN												
Chang 2011	HAI	(I) = 82.76 (J) = 72.41 (K) = 68.97	(I) = 78.15 (J) = 84.66 (K) = 86.16			(I) = 0.961 (J) = 0.954 (K) = 0.942		(I) = 0.850 (J) = 0.820 (K) = 0.791						
		ML: Symmetrical Margin SVM												
Cohen 2004	HAI	50.6	94.4			0.896								
		ML: SVM												
Cohen 2006	HAI	43	92			0.86	0.55							
		ML: Symmetrical Margin SVM												
Cohen 2008	HAI	50.6	94.4			0.896								
		ML: GTB optimized												
Ehrentraut 2018	HAI										93.7	79.7	85.7	
		ML: NLP												
Gerbier 2011	HAI										85.8	79.1		
		ML: Machine Learning												
Gomez-Vallejo 2016	HAI					0.702								0.62

Ref	Sensitivity	Specificity	PPV	NPV	Accuracy	AUROC	LR+	LR−	CWA	Recall	Precision	F1
	Control: LR											
Chen 2014	38.0	86.7	70.4	62.7		0.759	2.85	0.72				
	Control: medical scoring system											
Chang 2011	(I) = 68.97 (J) = 62.07 (K) = 68.97	(I) = 91.50 (J) = 92.59 (K) = 84.62			(I) = 0.912 (J) = 0.922 (K) = 0.844	(I) = 0.871 (J) = 0.830 (K) = 0.791						
	ML: Asymmetrical Margin SVM											
Cohen 2004	92	72.2			0.744							
	ML: AdaBoost											
Cohen 2006	45	95			0.86				0.58			
	ML: Asymmetrical Margin SVM											
Cohen 2008	92	72.2			0.744							
	ML: SVM optimized											
Ehrentraut 2018										89.8	83.1	84.8
Gerbier 2011												
Gomez-Vallejo 2016												

Please cite this article in press as: Scardoni A, et al. Artificial intelligence-based tools to control healthcare associated infections: A systematic review of the literature. J Infect Public Health (2020), https://doi.org/10.1016/j.jiph.2020.06.006

Table 4 (*Continued*)

Ref	Sensitivity	Specificity	Accuracy	AUROC	CWA	Sensitivity	Specificity	Accuracy	CWA	Sensitivity	Specificity	Accuracy	CWA
Chen 2014													
	Control: LR												
Chang 2011	(I) = 82.76 (J) = 75.86 (K) = 68.97	(I) = 80.90 (J) = 81.63 (K) = 86.16	(I) = 0.988 (J) = 0.985 (K) = 0.989	(I) = 0.870 (J) = 0.831 (K) = 0.792									
Cohen 2004													
	ML: C4.5					ML: Naive Bayes				ML: IB1			
Cohen 2006	28	95	0.88		0.45	57	88	0.85	0.65	19	96	0.88	0.38
Cohen 2008													
Ehrentraut 2018													
Gerbier 2011													
Gomez-Vallejo 2016													

Ref		Sensitivity	Specificity	PPV	NPV	Accuracy	AUROC	Recall	Precision	F1
		ML: NLP augmented algorithm								
Branch-Elliman 2015	CAUTI	65	99.6	54.2	99.7					
		ML: NLP								
Haas 2005	Nosocomial Pneumonia	71	95	7.9	99.8					
		ML: Xgboost								
Savin 2018	Healthcare-associated ventriculitis and meningitis			34	94		0.83	0.32	0.39	0.34
		ML: Xgboost								
Taylor 2018	UTI	80.0	84.7			0.837				

Ref	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	AUROC	F1
				Control: standard surveillance method				
Branch-Elliman 2015								
Haas 2005								
				Control: LR				
Savin 2018							0.81	0.28
	ML: Reduced Xgboost			Control: UTI diagnosis				
Taylor 2018	74.5	84.7	0.825	41.3	84.7	0.751		

(I) Variable Group1, (J) Variable Group3, (K) Variable Group5. (PPV) Positive Predictive Value, (NPV) Negative Predictive Value, (CWA) Mean class-weighted accuracy, (AUROC) Area under the Receiver Operating Characteristic, (LR+) positive likelihood ratio, (LR−) negative likelihood ratio, (F1) *F*-measure.

Performance measures' pooled data are reported by type of infections in Tables 2–4 and detailed in the sections below. In Tables 2–4 we report performance measures for each assessed ML-based model, performance measures of control models, grouping included studies by type of infections.

ML-based models for Central Line-associated Bloodstream Infections (CLABSI) surveillance

Two studies reported data on ML-based models applied to CLABSI surveillance in, respectively, hospital and ICU settings [30,39] (Table 2). Beeler et al. [30] compared a ML-random forest model with non-ML traditional logistic regression model, (AUROC of 0.87 and 0.79, respectively) and validated the best performing ML one to derive patients' personalized daily risk of CLABSI. Parreco et al. [39] compared the performance of three different ML-based models, demonstrating Gradient boosted trees-ML model to have the highest accuracy, precision, sensitivity and negative predictive value. Overall, all ML-based models tested in Parreco et al. [39] had: (i) high specificity and NPV, low sensitivity and PPV, and had lower AUROC as compared to Beeler et al. [30]. The application of Logistic Regression (LR) models was different in the two studies, in Beeler et al. [30] a static model was applied, instead in Parreco et al. [39] a machine learning model of LR was applied. Both studies concluded demonstrating the potential benefits of applying accurate ML-based models for CLABSI real time/early identification and risk prediction.

ML-based models for sepsis surveillance

We retrieved two papers reporting on ML-based models for sepsis' detection: one using retrospective data to test and validate them [32], while one assessing their impact on clinical outcomes in the context of an experimental study design [40] (Table 2). Desautels et al. [32] showed a ML-based classification system using a minimal set of clinical data to perform better than traditional scoring systems (Sequential Organ Failure Assessment – SOFA Score, quick SOFA – qSOFA, Modified Early Warning Score – MEWS, Systemic inflammatory response syndrome – SIRS Score, Simplified Acute Physiology Score – SAPS score), both at admission (AUROC = 0.88, APR = 0.595), and 1–4 h preceding sepsis onset (Table 2), performance holding higher even with 60% randomly missing data. A randomized controlled trial by Shimabukuro et al. [40] assessed the efficacy of an ML-based sepsis detection algorithm on reducing hospital length of stay (LOS) and mortality, as compared to traditional automated sepsis score surveillance, reporting – respectively – a 20.6% and 12.4% decrease (Table 2). Overall, both studies reported ML-based models to detect sepsis more accurately than traditional clinical scores, retrospectively for epidemiological and performance evaluation purposes, and prospectively for preventive real-time evaluations.

ML-based models for Clostridium difficile *infection (CDI)*

We retrieved two studies reporting on ML-models to predict *Clostridium difficile* infection in the US [33,38] (Table 2). Oh et al. [38] contrasted the idea of applying static non-ML prediction models for CDI across institutions but, instead, propose a ML approach to derive setting-specific ML-based risk models for CDI and report performance measures of two of them developed from retrospective health data analysis with AUROC of, respectively, 0.82 at the Massachusetts General Hospital, and 0.75 at the University of Michigan Hospital (Table 2). As authors comment, these ML-EHR-based CDI risk stratification models allow for earlier and more accurate identification of high-risk patients and better targeting of infection prevention strategies, with high specificity but high false positives (low positive predictive value, Table 2). Escobar et al. [33] compared different techniques, including machine learning models, to predict recurrent *Clostridium difficile* infection (Table 2); they selected three best performing models – of which one ML-based – but did not report any discrimination advantage, or better calibration or explanatory power, as compared to simple logistic regression (Table 2), leading authors to conclude that the use of ML models remains limited in CDI recurrence prediction.

ML-based models for Surgical Site Infections (SSI)

We retrieved eight studies reporting on the application of ML-based models to predict, control and assess Surgical Site Infections and their risk factors pre and post-surgery in European and US surgery departments [35,37,41,43,45,47,49,53]. Performance measures of different ML-based models are reported in Table 3.

Two studies compared ML-models with standard logistic regression models for SSI prediction in colorectal surgical patients [41,43] reporting higher performance of Bayesian network classifier using different set of variables ($p = 0.002$) in one study [41], while less supportive results were reported in the other [43] where ML-based models had higher performance than LR only prior to surgery [43] and only for specific ML classification methods (support vector machines, Table 3). Promising results of applying ML-based models to SSI prediction comes from neurosurgery [45], and head and neck surgery settings [47]. The use of Artificial neural networks (ANN) algorithms showed good results in predicting SSI in free flap reconstruction, performing better in postoperative prediction [47]; similarly, Natural Language Processing (NLP) detection approach showed the highest detection accuracy for SSI infections [45] (Table 3). A multi-center study assessed the SSI predictive value of ML-based models incorporating data from daily clinical wound assessment and reporting the best performing model to have moderate PPV (0.35) and high NPV (0.93) for identification of SSI in advance of clinical diagnosis [53]. Soguero-Ruiz et al. [49] explored different ML-based models (linear and non-linear SVM) for SSI temporal prediction fueled by blood test results and reported pre-operative and post-operative accuracy to range from, respectively, 0.69 and 0.67 to 0.91 and 0.90. Finally, a large US study developed and tested against the US National Surgical Quality Improvement project (NSQIP) data prototype ML-based systems to detect superficial, deep and organ/space SSI reporting them to have high specificity (0.78–0.98) and high NPV (>0.98) [35]. Overall, ML models' sensitivity ranged between 0.42 and 0.80, specificity ranged between 0.64 and 0.93, Positive Predictive Value between 0.33 and 0.53, Negative Predictive Value between 0.87 and 0.99, accuracy between 75.8 and 90 and AUROC between 0.436 and 0.896.

ML-based models for Healthcare Associated Infections (HAIs)

Eight included studies reported on ML-based models for the control of HAIs in general (Table 4), either to retrospectively identify HAIs' determinants and risk factors, or to prospectively predict their occurrence [7,28,29,44,46,48,51,52] (Table 4). Included studies compared different ML-based models or compared them with non ML models. Swedish data comparing two different ML-based models, namely SVMs and GTB showed the latter to perform better in terms of percent recall (93.7) and precision (79.7) [52]. Relatively high HAI detection performances of ML-based models were reported in the US [36], France (0.79 precision) [7] and Spain (0.70 precision) [51] (Table 4). Studies conducted in China [48] on lung cancer patients and Taiwan [46] compared ML-based models fed by Electronic Health Records data to LR and manual scoring models reporting high discrimination power of ANN with AUC ranging from 0.79 to 0.85 (Table 4), although not statistically higher than

LR in data from Taiwan [46]. Three overlapping studies conducted by Cohen et al. [28,29,44] investigated the performance of one-class support vector machines for HAI detection underlying how, as compared to two-class approach, they better accounted for imbalance in HAI data prevalence (i.e. few positive and a lot of negative cases) reaching 0.92 sensitivity, 0.72 specificity and 0.74 accuracy in best performing models (SVMs with asymmetrical margin [28]. In 2006 Cohen et al. [44] tested and compared several ML classifiers reporting sensitivity and specificity ranging from, respectively, 0.49 and 0.74 to 0.87 and 0.86 (Table 4). Overall, ML models' sensitivity ranged between 0.19 and 0.92, specificity ranged between 0.72 and 0.96 and accuracy between 0.70 and 0.96.

Single study HAIs

Of the 4 studies evaluating ML approaches on other HAI, two papers focused on urinary tract infections [31,42], one on healthcare associated ventriculitis and meningitis [50], one on nosocomial pneumonia [34] (Table 4). With regard to urinary tract infections, US data reported XGboost to perform best, as compared to other ML algorithms, as well as compared to provider judgment, antibiotic administration and documentation of UTI diagnosis [42] (Table 4), while data from a different study setting reported NLP-based models not to perform as well as standard surveillance methods [31]. Savin et al. [50] showed ML to be an effective approach to identify risk factors for healthcare-associated ventriculitis and meningitis with particular reference to Xgboost algorithms performing better than other assessed ML-based algorithms (Table 4). Retrospective analysis conducted in neonatal intensive care units produced performance data on ML-based automated surveillance system for nosocomial pneumonia (sensitivity: 0.71, specificity 0.99, PPV 0.08 and NPV >0.99) [34].

Discussion

Our review identified 27 studies in which ML-based models were applied to HAIs surveillance and control in different clinical settings. Overall, there is moderate evidence that ML-based models perform equal or better as compared to non-ML approaches and that they reach relatively high-performance standards. However, heterogeneity amongst the studies was very high and did not dissipate significantly in subgroup analyses, by type of infection or type of outcome. More than half of included studies were conducted in the US and the majority of studies focused on surgical site infections. Available comparative data are between different ML-based models and: clinical scores, standard or automated (rule-based) surveillance models and logistic regression statistical algorithms; 63% of studies had a predictive approach, while 37% had a retrospective approach to risks identification for HAIs.

Digitalization is revolutionizing "*the way humans create, exchange, and distribute value*" [51], and rapidly shaping all aspects of society, including healthcare [8,56,57]. In this context, there is no doubt that artificial intelligence tools' application to the different fields of medicine will dramatically improve diagnostics, treatments and ultimately health outcomes. The adoption of ML-based instruments in health has been taken up at different paces in different fields of medicine. As summarized in a recent review, the areas in which research on artificial intelligence is more advanced are: cancer, nervous system and cardiovascular diseases [58], with promising applications to, for example, cancer mutations' identification [59,60], cardiovascular events' prediction [55,61,62], among others. We have previously reviewed and pooled the application of ML in orthopedics reporting a still preliminary – although expanding – phase of ML adoption, mostly linked to the use of imaging data [54]. More in general, the arguments around the application of AI in health are mostly framed around the concept of clinical decision support, or better said, around the concept of supporting physicians in their tasks to take decisions "in the absence of certitude" [54] in clinical contexts. A public health perspective on artificial intelligence is less frequently adopted. We have recently argued that as public health in Europe and across the globe faces substantial challenges – including the burden of HAIs and associated rise of antimicrobial resistance [63] – we should seek to better understand the potential of artificial intelligence use in supporting public health efforts [38,57]. How can ML tools support emerging public health threats through preventive approaches? In the current paper we make the case of HAIs' prevention and control. From all the studies included in the present review -although largely heterogeneous – it clearly emerges how ML-based models would allow for earlier and more accurate identification of high-risk patients and better targeting of infection prevention strategies in healthcare facilities with ultimate decreased incidence and costs. Indeed, in a generalized context where hospitals are struggling to control expenses, despite quality improvements in healthcare, HAIs remain a major cost. A meta-analysis published on Jama estimated that in the US the total annual costs for the 5 major HAIs is $9.8 billion [8]. None of the studies retrieved in our review reported cost-effectiveness analysis on the application of ML-based intervention for HAIs surveillance and control, a research area worth exploring in the near future. More in general, the field of cost-effectiveness analysis of ML-based tools in healthcare is still poorly explored, some data support the cost-effectiveness of targeted screenings using a ML risk prediction algorithms [64] but solid evidence on cost-effectiveness of ML in the different areas of medicine is missing [1]. Almost all included studies provided performance measures of newly developed or adapted ML models while limited data is available on their validation [65] and on their implementation in clinical practice. In fact, we could retrieve only one RCT that reported data on the experimental use of a ML-based severe sepsis prediction system showing reduced average length of stay (−21%) and in-hospital mortality (−12%). In addition, the single-center study's relatively small sample size, the short study period limits the generalisability of its results. Overall, scant data is available on ML tools' use and impact in clinical practice, this confirming that despite ongoing lively discussion around the potential of artificial intelligence in support of healthcare delivery, evidence is still largely lacking. Our data demonstrate that research outputs are progressively accumulating on the topic, but we are still far from validation, adoption and scaling up of ML-based models for HAIs control and almost no data exist on their impact on clinical, organizational or economic outcomes. A number of pillars need to be strengthened to get to widespread adoption of ML-based tools for HAIs control; these relate to the availability of data and technical infrastructures, to the development and validation of highly performing models and to the tackling of the normative, cultural, behavioral and organizational determinants challenging their adoption. First, large volumes of electronic health data should be available, accessible and linkable so as to inform and fit machine learning algorithms. Indeed, the value of machine learning predictive algorithms is to unlock and make meaningful use of large, complex data. The availability of electronic health data varies widely across countries, regions and single healthcare facilities. In the US it is estimated that more than 90% of hospitals have an electronic medical system in place [10], in Europe the distribution of the Digital Health Index that assesses Countries' digital health readiness lists Estonia and Denmark as preforming best in Europe [67] while other countries lack far behind . Second, our data demonstrate that multidisciplinary teams are developing ML-based prediction models for HAI detection and control whose performance seem to outperform non ML-based models; however, wide heterogenicity in terms of: type of data feeding the models, studies' setting of implementation, type ML tested models and assessed performance measures make it difficult to derive compar-

Box 2: Key findings and significance for clinical practice and public health

Key findings

- Evidence is accumulating on the performance of different ML-based models for HAIs detection in different settings.
- There is moderate evidence that ML-based models perform equal or better as compared to non-ML approaches (i.e. clinical scores, standard or automated/rule-based surveillance models and logistic regression statistical algorithms).
- ML-based models performance metrics favor specificity and negative predictive value, more than sensitivity and other performance measures, thus underlining the potential of ML models to discriminate non-infected subjects.
- There is wide heterogeneity in study designs which prevent to derive comparative analysis and quantitatively pool available performance data.
- Available evidence is limited to ML-based models' performance assessment and still scant reporting of their application and impact on clinical practice is available.

Significance for clinical practice and public health

- In the near future ML-based HAIs' control systems might be integrated in hospital clinical practice.
- ML-based HAIs' control systems in the future might improve effectiveness and reduce costs of patient safety interventions.
- ML might lead to improved understanding of HAIs risk factors, improved patient risk stratification, as well as timely or real-time HAIs detection and control.
- For the implementation and use of ML-based HAIs' control systems in clinical practice large volumes of electronic health data should be available, accessible and linkable.
- Strengthened and multidisciplinary collaboration between IT and clinical disciplines is envisaged to promote the adoption and use of ML-based HAIs' control systems in clinical practice.

ative analysis and quantitatively pool available evidence. Overall, in line with what reported in the literature on automated surveillance systems [66], our data suggest ML-based models performance metrics favor specificity and negative predictive value, more than sensitivity and other performance measures, thus underlining the potential of ML models to detect the non-infected subjects. Third, despite an ongoing fruitful debate around the potential for adoption of ML-based solutions in healthcare, scant elements are available on normative, cultural, behavioral and organizational factors still hampering their adoption in infection control [68,69]. More in details, how technological innovation will change the designing and implementation of HAIs surveillance and how this will modify the roles and functions of clinical staff and the organization of health services delivery? Data informing such reasoning is still scant [66]; the transition from resource intensive conventional manual surveillance lacking standardization to semi-automated, automated and ML-based automated surveillance should be modulated on the basis on the aims and scale of surveillance, differentiating, for instance, between research purposes, in-hospital quality improvement, or national and international-level surveillance [66]. Despite potential advantages offered by ML-based tools for HAIs surveillance in terms of reduced costs, improved quality and efficiency, their introduction has to deal with, among others, low acceptance by the medical community and heterogeneity of hospital information systems hampering benchmarking [66,70].

Our review has both strengths and limits. To our knowledge this is the first systematic review on the application of artificial intelligence-based tools to control healthcare associated infections. The solid and rigorous methodology applied allowed us to retrieve and pool a comprehensive set of data which offer a complete overview of the state of the art of research and practice in this field. In addition, quality assessment of included studies proved them to be overall of good methodological quality. However, despite having meticulously extracted outcomes' data, heterogeneity amongst the studies prevented us from quantitatively pooling them in meta-analysis. Not only included studies focused on different HAIs, in different clinical settings, and applied a wide range of different performance measures, most importantly they developed and tested different ML models fitted with different data sources.

Our findings demonstrate that research outputs on how to apply ML-based solutions to HAIs surveillance are progressively accumulating; to date available evidence mainly focuses on the development and testing of detection and prediction models while their adoption and impact in clinical practice are still far from being explored or exploited. In the future efforts should be devoted on one hand to further develop and validate but also adopt, assess and scale up ML-based tools for HAIs control, and – on the other hand – to ensure the availability of accurate and reliable data stored in electronic health records that can inform, maintain and finetune their implementation (Box 2).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jiph.2020.06.006.

References

[1] Magill SS, O'Leary E, Janelle SJ, Thompson DL, Dumyati G, Nadle J, et al. Changes in prevalence of health care-associated infections in U.S. Hospitals. N Engl J Med 2018;379(18):1732–44.
[2] World Health Organization (WHO). The critical role of infection prevention and control. Health care without avoidable infections; 2016.
[3] Cassini A, Plachouras D, Eckmanns T, Abu Sin M, Blank HP, Ducomble T, et al. Burden of six healthcare-associated infections on European population health: estimating incidence-based disability-adjusted life years through a population prevalence-based modelling study. PLoS Med 2016;13(10):e1002150.
[4] Allegranzi B, Kilpatrick C, Storr J, Kelley E, Park BJ, Donaldson L. Global infection prevention and control priorities 2018–22: a call for action. Lancet Global Health 2017;5(12):e1178–80.
[5] Bagheri Nejad S, Allegranzi B, Syed SB, Ellis B, Pittet D. Health-care-associated infection in Africa: a systematic review. Bull World Health Organ 2011;89(10):757–65.
[6] Ling ML, Apisarnthanarak A, Madriaga G. The burden of healthcare-associated infections in Southeast Asia: a systematic literature review and meta-analysis. Clin Infect Dis 2015;60(11):1690–9.
[7] Gerbier S, Yarovaya O, Gicquel Q, Millet A-l, Smaldore V, Pagliaroli V, et al. Evaluation of natural language processing from emergency department computerized medical records for intra-hospital syndromic surveillance. BMC Med Inform Decis Mak 2011;11:50.
[8] Zimlichman E, Henderson D, Tamir O, Franz C, Song P, Yamin CK, et al. Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system. JAMA Intern Med 2013;173(22):2039–46.
[9] The Joint Commission. A complimentary publication of The Joint Commission Issue 58; 2017.
[10] Mitchell BG, Hall L, Halton K, MacBeth D, Gardner A. Time spent by infection control professionals undertaking healthcare associated infection surveillance: a multi-centred cross sectional study. Infect Dis Health 2016;21(1):36–40.
[11] Sips ME, Bonten MJM, van Mourik MSM. Automated surveillance of healthcare-associated infections: state of the art. Curr Opin Infect Dis 2017;30(4):425–31.

[12] Ke CM, Huang FJ, Lee SS, Chen YS, Hsieh PJ, Lin YE. Use of data mining surveillance system in real time detection and analysis for healthcare-associated infections. BMC Proc 2011;5(Suppl 6):P235.
[13] Halverson CA. Activity theory and distributed cognition: or what does CSCW need to do with theories? Comput Support Coop Work (CSCW) 2002;11(1–2):243–67.
[16] Fernández-Delgado M, Cernadas E, Barro S, Amorim D, Fernández-Delgado A. Do we need hundreds of classifiers to solve real world classification problems?; 2014.
[17] Cabitza F, Banfi G. Machine learning in laboratory medicine: waiting for the flood? Clin Chem Lab Med (CCLM) 2018;56(4):516–24.
[18] Brynjolfsson E, Mitchell T. What can machine learning do? Workforce implications. Science 2017;358(6370):1530–4.
[19] Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. N Engl J Med 2017;376(26):2507–9.
[20] Deo RC. Machine learning in medicine. Circulation 2015;132(20):1920–30.
[21] Fitzpatrick F, Doherty A, Lacey G. Using artificial intelligence in infection prevention. Curr Treat Opt Infect Dis 2020:1–10.
[22] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(4):e1312.
[23] Amerio A, Stubbs B, Odone A, Tonna M, Marchesi C, Nassir Ghaemi S. Bipolar I and II disorders: a systematic review and meta-analysis on differences in comorbid obsessive-compulsive disorder. Iran J Psychiatry Behav Sci 2016;10(3):e3604.
[24] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8(5):336–41.
[25] Smeulers M, Lucas C, Vermeulen H. Effectiveness of different nursing handover styles for ensuring continuity of information in hospitalised patients. Cochrane Database Syst Rev 2014:6.
[26] Royce S, Falzon D, van Weezenbeek C, Dara M, Hyder K, Hopewell P, et al. Multidrug resistance in new tuberculosis patients: burden and implications. Int J Tuberc Lung Dis 2013;17(4):511–3.
[27] Skrahina A, Hurevich H, Zalutskaya A, Sahalchyk E, Astrauko A, Hoffner S, et al. Multidrug-resistant tuberculosis in Belarus: the size of the problem and associated risk factors. Bull World Health Organ 2013;91(1):36–45.
[28] Cohen G, Hilario M, Sax H, Hugonnet S, Pellegrini C, Geissbuhler A. An application of one-class support vector machine to nosocomial infection detection. Stud Health Technol Inform 2004;107(Pt 1):716–20.
[29] Cohen G, Sax H, Geissbuhler A. Novelty detection using one-class Parzen density estimator. An application to surveillance of nosocomial infections. Stud Health Technol Inform 2008;136:21–6.
[30] Beeler C, Dbeibo L, Mph KK, Thatcher L, Webb D, Bah A, et al. Machine learning. AJIC: Am J Infect Control 2018;46(9):986–91.
[31] Branch-Elliman W, Strymish J, Kudesia V, Rosen AK, Gupta K. Natural language processing for real-time catheter-associated urinary tract infection surveillance: results of a pilot implementation trial. Infect Control Hosp Epidemiol 2015;36(9):1004–10.
[32] Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shimabukuro D, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach corresponding author. JMIR Med Inform 2016;4(3):e28.
[33] Escobar GJ, Baker JM, Kipnis P, Greene JD, Mast TC, Gupta SB, et al. Prediction of recurrent *Clostridium difficile* infection using comprehensive electronic medical records in an integrated healthcare delivery system. Infect Control Hosp Epidemiol 2017;38(10):1196–203.
[34] Haas JP, Mendonc EA. Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients. Am J Infect Control 2005;33(8):439–43.
[35] Hu Z, Simon GJ, Arsoniadis EG, Wang Y, Kwaan MR, Melton GB. Automated detection of postoperative surgical site infections using supervised methods with electronic health record data. Stud Health Technol Inform 2015;216:706–10.
[36] Hu Z, Melton GB, Moeller ND, Arsoniadis EG, Wang Y, Kwaan MR, et al. Accelerating chart review using automated methods on electronic health record data for postoperative complications. AMIA Annu Symp Proc 2016:1822–31.
[37] Ke C, Jin Y, Evans H, Lober B, Qian X, Liu J, et al. Prognostics of surgical site infections using dynamic health data. J Biomed Inform 2017;65:22–33.
[38] Oh J, Makar M, Fusco C, McCaffrey R, Rao K, Ryan EE, et al. A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. Infect Control Hosp Epidemiol 2018;39(4):425–33.
[39] Parreco JP, Hidalgo AE, Badilla AD, Ilyas O, Rattan R. Predicting central line-associated bloodstream infections and mortality using supervised machine learning. J Crit Care 2018;45:156–62.
[40] Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Resp Res 2017;4:e000234.
[41] Sohn S, Larson DW, Habermann EB, Naessens JM, Alabbad JY, Liu H. Detection of clinically important colorectal surgical site infection using Bayesian network. J Surg Res 2017;209:168–73.
[42] Taylor RA, Moore CL, Cheung K-H, Brandt C. Predicting urinary tract infections in the emergency department with machine learning. PLoS ONE 2018;13(3):e0194085.
[43] Weller GB, Lovely J, Larson DW, Earnshaw BA, Huebner M. Leveraging electronic health records for predictive modeling of post-surgical complications. Stat Methods Med Res 2017;27(11):3271–85.
[44] Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. Artif Intell Med 2006;37:7–18.
[45] Campillo-gimenez B, Garcelon N, Jarno P, Chapplain MJ, Cuggia M. Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France. Stud Health Technol Inform 2013;192:572–5.
[46] Chang Y-J, Yeh M-L, Li Y-C, Hsu C-Y, Lin C-C. Predicting hospital-acquired infections by scoring system with simple parameters. PLoS ONE 2011;6(8):e23137.
[47] Kuo P-J, Wu S-C, Chien P-C, Chang S-S, Hsieh Y, Hsieh C-H. Artificial neural network approach to predict surgical site infection after free-flap reconstruction in patients receiving surgery for head and neck cancer. Oncotarget 2018;9(17):13768–82.
[48] Chen J, Pan Q-S, Hong W-D, Pan J, Zhang W-H, Xu G, et al. Use of an artificial neural network to predict risk factors of nosocomial infection in lung cancer patients. Asian Pac J Cancer Prev 2014;15(13):5349–53.
[49] Soguero-Ruiz C, Fei Wang ME, Jenssen R, Augestad M, Rojo Alvarez J-L, Mora Jiménez I, et al. Data-driven temporal prediction of surgical site infection. AMIA Annu Symp Proc 2015:1164–73.
[50] Savin I, Ershova K, Kurdyumova N, Ershova O, Khomenko O, Danilov G, et al. Healthcare-associated ventriculitis and meningitis in a neuro-ICU: incidence and risk factors selected by machine learning approach. J Crit Care 2018;45:95–104.
[51] Gomez-Vallejo HJ, Uriel-Latorre B, Sande-Meijide M, Villamarin-Bello B, Pavon R, Fdez-Riverola F, et al. A case-based reasoning system for aiding detection and classification of nosocomial infections. Decis Support Syst 2016;84(C):104–16.
[52] Ehrentraut C, Ekholm M, Tanushi H, Tiedemann J, Dalianis H. Detecting hospital-acquired infections: a document classification approach using support vector machines and gradient tree boosting. Health Inform J 2018;24(1):24–42.
[53] Sanger PC, Ramshorst GHV, Mercan E, Huang S, Hartzler A, Armstrong CA, et al. A prognostic model of surgical site infection using daily clinical wound assessment. J Am Coll Surg 2016;223(2):259–70.
[54] Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. Front Bioeng Biotechnol 2018;6:75.
[55] Zhao J, Feng Q, Wu P, Lupu RA, Wilke RA, Wells QS, et al. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. Sci Rep 2019;9(1):717.
[56] Azzopardi-Muscat N, Ricciardi W, Odone A, Buttigieg S, Zeegers Paget D. Digitalization: potentials and pitfalls from a public health perspective. Eur J Public Health 2019;29(Suppl. 3):1–2.
[57] Odone A, Buttigieg S, Ricciardi W, Azzopardi-Muscat N, Staines A. Public health digitalization in Europe. Eur J Public Health 2019;29(Suppl. 3):28–35.
[58] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230–43.
[59] Kann BH, Thompson R, Thomas Jr CR, Dicker A, Aneja S. Artificial intelligence in oncology: current applications and future directions. Oncology (Williston Park, NY) 2019;33(2):46–53.
[60] Wood DE, White JR, Georgiadis A, Van Emburgh B, Parpart-Li S, Mitchell J, et al. A machine learning approach for somatic mutation discovery. Sci Transl Med 2018;10(457).
[61] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE 2017;12(4):e0174944.
[62] Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and machine learning for heart failure survival analysis. Stud Health Technol Inform 2015;216:40–4.
[63] Increasing appropriate vaccination: immunization information systems. In: Force CPST, editor. 2010.
[64] Hill NR, Sandler B, Mokgokong R, Lister S, Ward T, Boyce R, et al. Cost-effectiveness of targeted screening for the identification of patients with atrial fibrillation: evaluation of a machine learning risk prediction algorithm. J Med Econ 2020:1–8.
[65] Kakarmath S, Golas S, Felsted J, Kvedar J, Jethwani K, Agboola S. Validating a machine learning algorithm to predict 30-day re-admissions in patients with heart failure: protocol for a prospective cohort study. JMIR Res Protoc 2018;7(9):e176.
[66] van Mourik MSM, Perencevich EN, Gastmeier P, Bonten MJM. Designing surveillance of healthcare-associated infections in the era of automation and reporting mandates. Clin Infect Dis 2018;66(6):970–6.
[67] Baguelin M, Flasche S, Camacho A, Demiris N, Miller E, Edmunds WJ. Assessing optimal target populations for influenza vaccination programmes: an evidence synthesis and modelling study. PLoS Med 2013;10(10):e1001527.
[68] Pesapane F, Volonte C, Codari M, Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging 2018;9(5):745–53.
[69] Rawson TM, Ahmad R, Toumazou C, Georgiou P, Holmes AH. Artificial intelligence can improve decision-making in infection management. Nat Hum Behav 2019;3(6):543–5.
[70] Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016;279(2):329–43.