PLOS ONE

RESEARCH ARTICLE

Automated detection of hospital outbreaks: A systematic review of methods

Brice Leclère[1,2]*, David L. Buckeridge[3], Pierre-Yves Boëlle[4☯], Pascal Astagneau[5,6☯], Didier Lepelletier[2,7]

1 Department of Medical Evaluation and Epidemiology, Nantes University Hospital, Nantes, France, **2** MiHAR laboratory, Nantes University, Nantes, France, **3** Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, **4** UMR S 1136, Pierre Louis Institute of Epidemiology and Public Health, Pierre and Marie Curie University, Paris, France, **5** Department of Public Health, Pierre and Marie Curie University, Paris, France, **6** Centre de Coordination de la Lutte contre les Infections Nosocomiales Paris-Nord, Hôpital Broussais, Paris, France, **7** Department of Microbiology and Infection Control, Nantes University Hospital, Nantes, France

☯ These authors contributed equally to this work.
* brice.leclere@univ-nantes.fr

OPEN ACCESS

Citation: Leclère B, Buckeridge DL, Boëlle P-Y, Astagneau P, Lepelletier D (2017) Automated detection of hospital outbreaks: A systematic review of methods. PLoS ONE 12(4): e0176438. https://doi.org/10.1371/journal.pone.0176438

Editor: Andre Scherag, University Hospital Jena, GERMANY

Received: December 15, 2016

Accepted: April 10, 2017

Published: April 25, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Objectives

Several automated algorithms for epidemiological surveillance in hospitals have been proposed. However, the usefulness of these methods to detect nosocomial outbreaks remains unclear. The goal of this review was to describe outbreak detection algorithms that have been tested within hospitals, consider how they were evaluated, and synthesize their results.

Methods

We developed a search query using keywords associated with hospital outbreak detection and searched the MEDLINE database. To ensure the highest sensitivity, no limitations were initially imposed on publication languages and dates, although we subsequently excluded studies published before 2000. Every study that described a method to detect outbreaks within hospitals was included, without any exclusion based on study design. Additional studies were identified through citations in retrieved studies.

Results

Twenty-nine studies were included. The detection algorithms were grouped into 5 categories: simple thresholds (n = 6), statistical process control (n = 12), scan statistics (n = 6), traditional statistical models (n = 6), and data mining methods (n = 4). The evaluation of the algorithms was often solely descriptive (n = 15), but more complex epidemiological criteria were also investigated (n = 10). The performance measures varied widely between studies: e.g., the sensitivity of an algorithm in a real world setting could vary between 17 and 100%.

Conclusion

Even if outbreak detection algorithms are useful complementary tools for traditional surveillance, the heterogeneity in results among published studies does not support quantitative synthesis of their performance. A standardized framework should be followed when evaluating outbreak detection methods to allow comparison of algorithms across studies and synthesis of results.

Introduction

Hospital information systems are goldmines for infection preventionists and epidemiologists. The large amount of data that they contain can help to detect adverse events, highlight risk factors, and evaluate the effectiveness of preventive actions [1]. These big data differ substantially from the ones that epidemiologists traditionally handle, but thanks to innovative methods borrowed from machine learning, data mining and natural language processing, they can be used to improve the quality and safety of healthcare [2]. Indeed, recent literature reviews have shown how these methods have been successfully applied to identify nosocomial infections [1,3], adverse drug events [4], and a wide range of other complications within hospitals [5].

Identifying nosocomial infections is useful to detect hospital outbreaks, which, given the potential morbidity, disorganization and cost that they can cause, represent a menace to patients, caregivers and healthcare systems [6,7]. However, case identification is only the first step in the surveillance process, and epidemiologists must then search for patterns that substantiate epidemic spread [8].

Fortunately, a wide range of automated outbreak detection methods is available and routinely used for community syndromic surveillance. Several infection control teams have also studied the usefulness of these methods at the scale of their own hospital, but the results were heterogeneous, precluding straightforward conclusions. The objective of our study was to clarify this issue by summarizing the existing literature on hospital outbreak detection algorithms, and especially by describing the evaluations approaches and the detection performances when applicable.

Methods

Study selection

In order to give the most accurate summary of the literature, we followed a systematic literature review protocol. The search query was built as a union of three sets of terms that related to the following key words: hospital, outbreak and detection (See S1 Appendix, protocol not accessible).

The MEDLINE bibliographic database was searched using the PubMed search engine in April 2016. To ensure the highest sensitivity, no limitations were imposed on publication dates and languages. The results of the query were screened based successively on the title, the abstract and the full text. Every reference that described a method used to detect outbreaks within hospitals was included, without any exclusion based on study design. The references that related to community outbreak detection or national/regional nosocomial infection surveillance were not included. The citations of every included document were also screened in search of relevant additional references, a method called *snowballing*. Complementarily, we

performed *reverse snowballing* by identifying relevant documents that cited the included studies, using the Google Scholar search engine.

One author (BL) extracted data from the included studies using a standardized form. The variables of interest were the following: date of publication, country, study period, spatial and temporal scopes of the surveillance, events of interest, data sources, detection algorithms and evaluation methodology.

Data analysis

To classify the studies according to their methodology, we used a framework developed by Watkins et al. for early outbreak detection evaluation [9]. According to this framework, four types of evaluation approaches can be identified: descriptive, derived, epidemiological, and simulation-based. The descriptive approach does not rely on detection performance measures, but rather on the description of detected events (frequency, duration, etc.). The "derived" approach uses the results of a statistical model as a reference standard to evaluate detection methods. The epidemiological approach uses more complex definitions based on multifactorial and flexible methods such as expert judgment. The last approach is the use of simulations, i.e. synthetic data. It allows for a complete control of outbreak features, but the validity of the estimations in the real world is not guaranteed. Besides classification, the methodologies were also analyzed to determine the risk of specific biases [10].

The performance measures (sensitivity, specificity, positive and negative predictive values) of the detection algorithms were also extracted, along with their 95% confidence intervals. If the confidence intervals were not available, we computed them based on the available data. The inter-algorithm heterogeneity was measured using the I^2 statistics, which represents the percentage of variability that is due to heterogeneity between algorithms. As several studies used different algorithms, we also calculated an R^2 statistic to estimate the proportion of the inter-algorithm heterogeneity that was due to differences between studies. These R^2 were estimated using mixed-effect meta-regressions that included a study effect. All the statistical analyses were done using the R software version 3.2.0 and the metafor package.

The present article was prepared using the PRISMA checklist for items reporting in systematic reviews [11] (S2 Appendix).

Results

Twenty-nine studies were included at the end of the selection process (Fig 1). They are described in details in Table 1. In the next sections, we will describe these studies with regards to the type of surveillance in which the algorithms were used and to the methods on which these algorithms relied. Finally, we will examine the observed performances for each evaluation approach.

Surveillance scope

Across the studies, the algorithms were used to detect different types of events. Three studies aimed to detect every nosocomial outbreaks, without any additional precision regarding their size, duration or type [22,24,31]. In two other studies, the events corresponded to cases with a clinical definition, i.e. nosocomial invasive aspergillosis [21] and bloodstream infections [39]. The remaining studies focused on infections caused by specific organisms such as multidrug resistant (MDR) bacteria or organisms known to cause nosocomial infections. Additional data allowed some algorithms to be stratified by infection site (bloodstream, urinary tract, etc.), organism or resistance pattern [13–15,19,23,28,29,31,32,34,35,37].

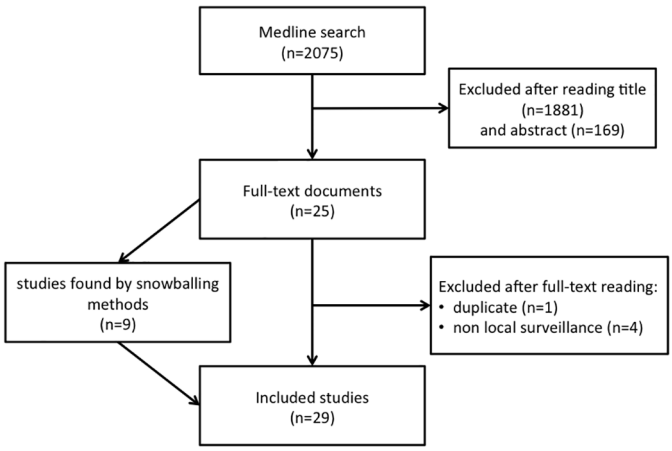

Fig 1. Study selection flow diagram.

https://doi.org/10.1371/journal.pone.0176438.g001

In most of the included studies (n = 24), the surveillance was implemented at the level of an entire hospital, but larger and smaller scopes were also reported. One study was conducted in a health trust consisting of 10 hospitals [31], and another one examined hospital-level outbreak detection based on the national surveillance data in England [23]. Conversely, in three studies, outbreaks were monitored at the level of a single intensive care unit [12,16,22]. Additional data allowed in six studies to stratify outbreak detection at different spatial levels, from the whole hospital to services and units [25–27,33–35].

Nearly every algorithm relied on either bacteriological laboratory results (n = 17) or nosocomial infection counts estimated by active surveillance from the infection control team (n = 10). Two studies additionally extracted admission-discharge-transfer data to provide a denominator for computing incidence rates [27,33]. Kikuchi et al. tested a more syndromic approach: instead of positive cases, their algorithm relied on counts of new occurrences of symptoms found in the electronic medical records [40].

The periods covered by these data varied between studies from 3 to 120 months, with a median of 24 months (inter-quartile range: 12.25–58.75).

Detection algorithms

Many different algorithms were implemented in the included studies, but they could all be classified into five categories: simple thresholds, statistical process control (SPC), statistical modeling, scan statistics and data mining methods. The trends of these categories over time are depicted in Fig 2.

With simple thresholds, an alert is triggered when the number of cases exceeded a threshold over which the number of infections in a given space and time is considered abnormal. These simple rules were used in six studies [24,28,32,34,38,39] and could either be an arbitrary threshold chosen by experts (e.g. three infections in two weeks in the same ward in the study by Huang et al. [27]) or a simple mathematical rule (e.g. a doubling of the ward's average monthly culture rate in the study by Schifman and Palmer [28]).

Algorithms based on SPC were the most commonly used in the included studies (n = 12). For these algorithms, the alert threshold is not defined arbitrarily but based on statistical variations of cases frequency in the past. SPC offers the possibility to monitor different types of statistical parameter, such as incidence count or rate [14,16,17,22,28,32,37], cumulative sums (CuSums) [21,22,30,35] or moving averages [30,35,37].

Table 1. Description of the included studies.

Study	Country	spatial scope of surveillance	spatial stratification	time unit of detection frequency	monitored types of infection	monitored measures	data sources	type of detection algorithms	type of evaluation	length of evaluation in month
Childress and Childress (1981) [12]	USA	intensive care unit of a university hospital	not applicable	month	*Serratia marcescens* infections	number of isolates	bacteriological lab results	SPC (thresholds based on endemic rate)	descriptive	12
Dessau and Steenberg (1993) [13]	Denmark	university hospital	none	week	organism specific infections	number of isolates	microbiology lab results	statistical modeling (time series analysis)	descriptive	12
Mylotte (1996) [14]	USA	university long term care facility (120 beds)	none	month	location-specific nosocomial infections	number of cases	ICP surveillance	SPC (thresholds based on endemic rate)	descriptive	96
Brossette et al. (1998) [15]	USA	university hospital	none	month	*Pseudomonas aeruginosa* infections	proportion of cases	bacteriological lab results and patient demographics	data mining (association rules)	descriptive	12
Arantes et al. (2003) [16]	Brazil	pediatric intensive care unit of a university hospital	not applicable	month	nosocomial infections	incidence rate of cases	IC surveillance	SPC (u-chart)	descriptive	36
Sagel et al. (2004) [17]	Germany	tertiary-care hospital (900 beds)	none	week	MRSA infections	number of isolates	IC surveillance	SPC (c-chart)	descriptive	12
Pentland et al. (2006) [18]	USA	university hospital	none	day	MDR-GN infections	number of isolates	bacteriological lab results	scan statistics	descriptive	24
Lamma et al. (2006) [19,20]	Italy	university hospital	wards	week	organism specific infections	number of cases	bacteriological lab results	statistical modeling (time series analysis)	descriptive	
Menotti et al. (2010) [21]	France	university hospital	none	month	nosocomial invasive aspergillosis	number of cases	IC surveillance	SPC (CuSum, LC-CuSum)	descriptive	24
Gomes et al. (2011) [22]	Brazil	intensive care unit of a university hospital	none	week	nosocomial infections	number of cases	IC surveillance	SPC (CuSum, u-chart, EWMA)	descriptive	24
Freeman et al. (2013) [23]	England	hospitals participating in national surveillance	none	week	12 species-specific infections 7 MDRO infections	number of cases	national IC surveillance	statistical modeling (quasi-Poisson model) and SPC (CuSum and)	descriptive	36

(*Continued*)

Table 1. (*Continued*)

Study	Country	spatial scope of surveillance	spatial stratification	time unit of detection frequency	monitored types of infection	monitored measures	data sources	type of detection algorithms	type of evaluation	length of evaluation in month
Du et al. (2014) [24]	China	tertiary-care hospital (3500 beds)	wards	day	nosocomial infections	number of isolates, diarrhea cases or surgical site infections	automated nosocomial infection surveillance	simple thresholds ($\geq$2 to 3 cases in 1 to 21 weeks)	descriptive	48
Faires et al. (2014)A [25]	Canada	community hospital (350 beds)	hospital, services and wards	day	*Clostridium difficile* infections	number of isolates	bacteriological lab results	scan statistics	descriptive	55
Faires et al. (2014)B [26]	Canada	community hospital (350 beds)	hospital, services and wards	day	MRSA infections	number of isolates	bacteriological lab results	scan statistics	descriptive	55
Lefebvre et al. (2015) [27]	France	2 university hospitals (1200 and 1800 beds)	Hospital and units	day	*Pseudomonas aeruginosa* infections	Number and incidence rate of isolates	bacteriological lab results	scan statistics	descriptive	112 and 78 (depending on the hospital)
Schifman and Palmer (1984) [28]	USA	university hospital (325 beds)	ward	month	organism and location specific infections	number of cases	ICP surveillance	simple thresholds ($\geq$ 2 times the average culture rate)	epidemiological	6
Brossette et al. (2000) [29]	USA	university hospital	unit	month	organism, location and antibiotic resistance specific infections	proportion of isolates	bacteriological lab results	Data mining (association rules)	epidemiological	15
Brown et al. (2002) [30]	USA	tertiary-care pediatric facility (330 beds)	not applicable	isolate	MRSA and VRE infections	number of isolates	bacteriological lab results	SPC (CuSum, moving average)	epidemiological	69
Ma et al. (2003) [31]	USA	10 hospitals of a university medical center	unit	month	organism, location and antibiotic resistance specific infections	number of isolates	bacteriological lab results	Data mining (association rules)	epidemiological	3
Hacek et al. (2004) [32]	USA	university hospital (688 beds)	none	month	organism specific infections	number of isolates and incidence rate of isolates	bacteriological lab results	simple thresholds (100% increase in 2 month, $\geq$50% increase in 3 months) and SPC (Shewart chart)	epidemiological	96
Wright et al. (2004) [33]	USA	university hospital (656 beds)	hospital, service and ward	week	location, organism, type and resistance specific infections	number of isolates	bacteriological lab results and admission-discharge-transfer	SPC (user-definable control charts)	epidemiological	13

(*Continued*)

Table 1. (*Continued*)

Study	Country	spatial scope of surveillance	spatial stratification	time unit of detection frequency	monitored types of infection	monitored measures	data sources	type of detection algorithms	type of evaluation	length of evaluation in month
Huang et al. (2010) [34]	USA	university hospital (750 beds)	hospital, services and wards	day	31 organism specific infections	number of isolates	bacteriological lab results	scan statistics	epidemiological	60
Carnevale et al. (2011) [35]	USA	general and pediatric hospital (800 beds)	hospital and units	day	organism specific infections	number of isolates	bacteriological lab results	SPC (CuSum, EWMA), scan statistics, data mining (WSARE)	epidemiological	24
Nishiura (2012) [36]	Japan	not implemented	none	month	-	number of isolates	IC surveillance	statistical modeling (Poisson model)	epidemiological	-
Tseng et al. (2012) [37]	Taiwan	university hospital (2200 beds)	none	week	MDR organism infections	number of isolates	bacteriological lab results	SPC (control charts ± hierarchical clustering)	epidemiological	12
Mellmann et al. (2006) [38]	Germany	university hospital (1480 beds)	wards	week	MRSA infections	number of isolates	bacteriological lab results	simple thresholds (2 isolates in 2 weeks, ± molecular typing)	derived	60
Charvat et al. (2009) [39]	France	university hospital (878 beds)	none	day	bloodstream infections	number of cases	IC surveillance	simple thresholds (delay between cases)	derived	120
Kikuchi et al. (2007) [40]	Japan	prefectoral central hospital	wards	day	symptoms	number of cases	electronic medical records (symptoms)	statistical modeling (linear model)	simulation	15
Skipper. (2009) [41]	Danemark	university hospital	none	day	simulated	number of isolates	bacteriological lab results	statistical modeling (Poisson model)	simulation	

MRSA: methicillin resistant *Staphylococcus aureus*, VRE: vancomycin resistant *Enterococcus*, IC: infection control, MDR: multi-drug resistant, GN: Gram negative, SPC: statistical process control, EWMA: exponentially-weighted moving average, WSARE: 'What's Strange About Recent Events?' algorithm, (LC-)CuSum: (Learning curve) cumulative sums.

https://doi.org/10.1371/journal.pone.0176438.t001

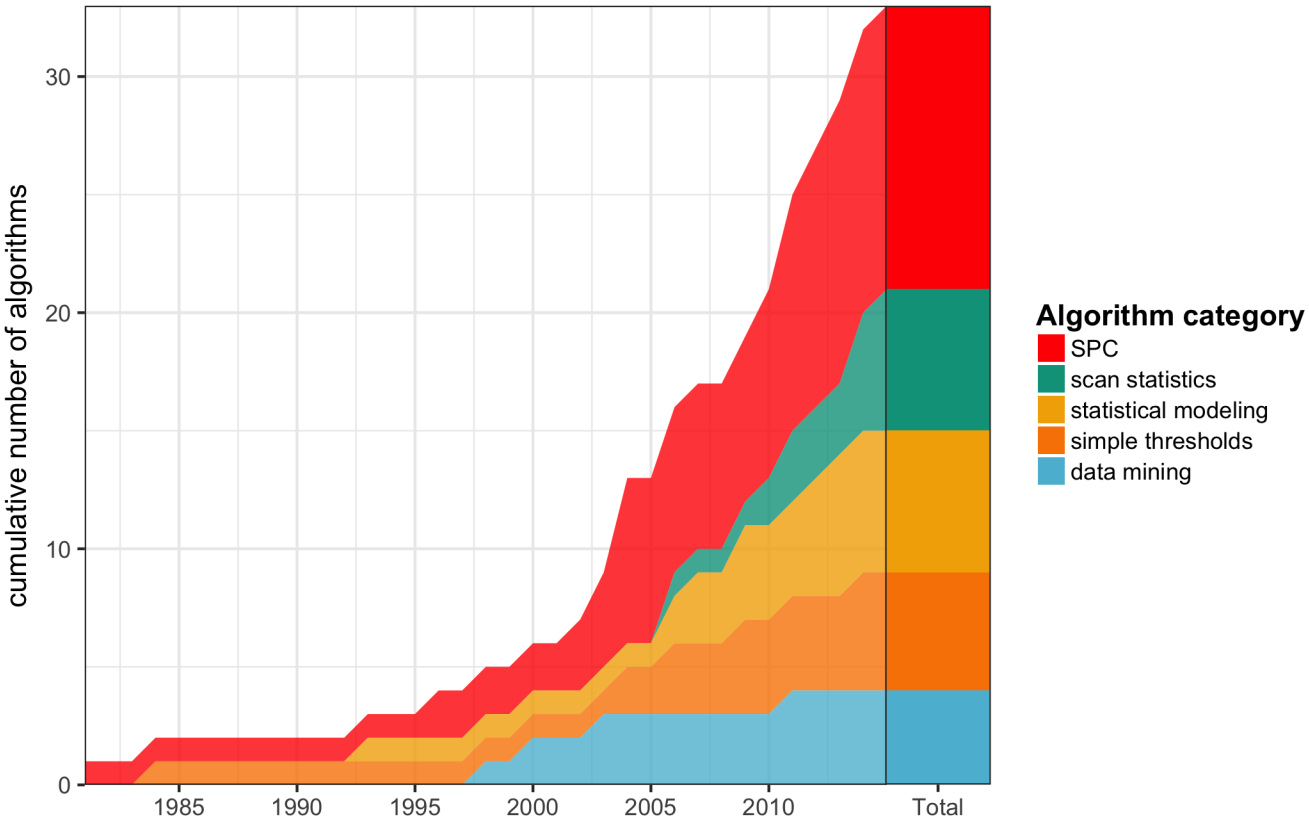

Fig 2. Cumulative count of detection algorithms found in the literature over time, by category. SPC: statistical process control.

https://doi.org/10.1371/journal.pone.0176438.g002

Statistical models were used in six studies [13,19,23,36,40,41]. They mostly consisted of multivariate Poisson regressions that allowed taking into account predictable factors of fluctuation in the number of infection cases, such as seasonality.

Elaborating on these models, scan statistics represented another popular category of algorithms (n = 6) [18,25–27,34,35]. It even served as a reference standard in an additional study by Mellmann et al. [38]. Because they use adjustable surveillance windows, they are more flexible than traditional statistical modeling and can detect events at different space and time scales.

Data mining methods constituted the last category of algorithms, which was used in four studies. Three related studies used association rules to automatically detect interesting changes in infection occurrence and resistance patterns [15,29,31]. A third tested an algorithm called 'What's Strange About Recent Events?' (WSARE) [35], which relies on Bayesian networks and associations rules [42].

Evaluation results

Descriptive approach. Fifteen of the included studies provided a descriptive evaluation of the algorithms' results [12–19,21–27]. All of them showed that detection algorithms implemented in real hospital settings were able to generate relevant alerts, and some of them reported interesting additional observations. For example, Gomes et al. noted a complementarity between SPC algorithms: Shewart charts were better for detecting large deviations from

the mean number of infections, while CuSums and exponentially-weighted moving averages were more suitable for detecting smaller deviations [22].

Freeman et al. noted that adding data about antibiotic resistance more than doubled the overall number of alerts generated by a Poisson-based model [23].

More recently, two studies by Faires et al. provided interesting insights about how a scan statistic-based algorithm compared to traditional infection control surveillance: it retrospectively identified most of the outbreaks already investigated by the infection control team but also flagged other potentially undetected epidemic events [25,26].

Epidemiological approach. The epidemiological approach was the second most frequently used evaluation design (n = 10). Its implementation, however, differed quite noticeably between studies. Some of them relied on the judgment of one [28,29,31] or several experts [30,33–35] to classify the alerts while others compared them to a list of previously identified outbreaks [36,37]. A last one used molecular typing, a common method for confirming clonal outbreaks, i.e. infections caused by the same strain [32].

Experts' evaluation of the alerts allowed the computation of positive predictive values (PPVs). As PPVs depend on the prevalence of the outbreaks, it was difficult to compare them across different surveillance settings, but they were overall superior to 75%, reaching a maximum at 96.5% for the CuSum algorithm in the study by Brown et al. [30]. Additionally, Carnevale et al. [35] combined multiple sources of alert to estimate the overall number of true positives. This allowed the estimation of sensitivity measures, which varied from 21 to 31% (Fig 3).

Out of the four studies that relied on a panel of experts, three reported inter-rater agreement estimated by Cohen's kappa. Using binary classifications, Wright et al. [33] and Huang et al. [34] reported good agreement ($\kappa = 0.82$ and 0.76 respectively) whereas Carnevale et al [35] reported lower results (from 0.11 to 0.49 on multiple pairs of raters).

Two studies provided estimates of the four traditional performance measures by comparing their algorithms to an "epidemiological" reference standard. Hacek et al. [32] used molecular typing, in a subset of potential alerts selected by the infection control team. Using this method, they reported that traditional surveillance was less specific than simple thresholds and SPC methods, with comparable sensitivity levels (Fig 3).

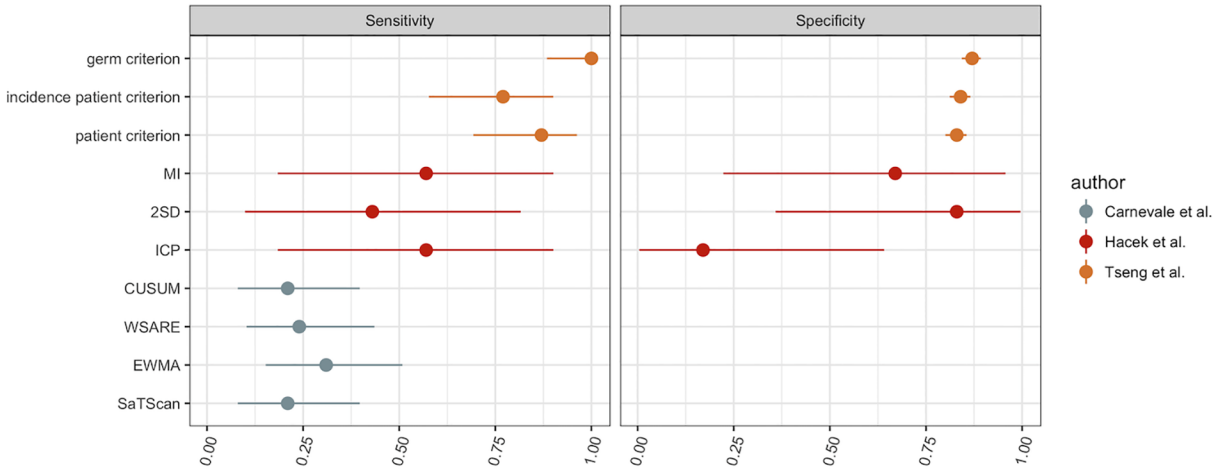

Fig 3. Sensitivity and specificity of the detection algorithms evaluated with the epidemiological approach (with 95% confidence intervals). Patient criterion: control chart based on the number of infected patients; incidence patient criterion: control chart based on the incidence of infected patients; germ criterion: control chart based on the number of positive results; MI: monthly increase; ICP: infection control surveillance; 2SD: control chart based on the number of positive results; WSARE: What's Strange About Recent Events?; SaTScan: scan statistics; EWMA: Exponentially-Weighted Moving Average; CUSUM: Cumulative sum.

https://doi.org/10.1371/journal.pone.0176438.g003

Alternatively, Tseng et al. [37] evaluated several SPC methods in comparison to the traditional infection control surveillance of vancomycin-resistant enterococcal infections. With the best parameters, the sensitivity and specificity of the algorithms ranged respectively from 77 to 100% and from 83 to 87% (Fig 3).

For these epidemiological approaches, we were able to estimate the inter-algorithm heterogeneity for sensitivity and specificity: according to the I^2 statistic, this heterogeneity accounted for respectively 83.5% and 67.4% of the overall variability. In the meta-regressions, the study effect explained respectively 100% and 45.33% of this heterogeneity (R^2 statistics).

Derived and simulation approaches. With the 'derived' and 'simulation' approaches, outbreaks are either statistically defined or identified in advance in the simulated datasets. The detection performances are thus more easily calculated and can even be estimated for different situations by modifying the parameters of the statistical definitions or the simulated datasets. As a result, the studies that used these approaches often reported ranges of performance measures as opposed to point estimates. For example, Kikuchi et al. [40] reported sensitivity and specificity measures for a linear model varying from 80 to 100% and 10 to 95% respectively, depending on the attack rates of the simulated outbreaks, while Skipper [41] reported sensitivity varying from 0 to 100% depending on the type of outbreak simulated, and on the parameters of their multistate Poisson model.

The derived approach also provided a straightforward reference standard for Mellmann et al. [38] to compare different detection methods. They estimated that traditional surveillance was more specific (97.3% vs. 47.3%) but less sensitive (62.1% vs. 100%) than a simple rule based on the frequency of cases. A rule based on both the frequency of cases and the results of molecular typing gave the best overall performance with a sensitivity of 100% and a specificity of 95.2%.

Discussion

Our literature review yielded 29 studies published between 1981 and 2015 that described algorithms for automated detection of nosocomial outbreaks. Among the different types of algorithms, those that were based on SPC were the most commonly used over the reviewed period. They have been applied for a long time in various fields of quality control and have been recommended for infection control surveillance for about two decades [43]. In the most recent studies, however, scan statistics have been used increasingly, as the popularity of the SaTScan package [44] rose in epidemiological research.

The surveillance scopes and settings in which these algorithms were implemented as well as the methods used to evaluate their performance varied quite noticeably between studies. According to our meta-regressions, the differences between studies explained a large part of the heterogeneity between of the results of this review. This heterogeneity did not allow us to estimate pooled accuracy measures using meta-analysis, and also precluded the comparison of the different algorithm categories. We acknowledge that our literature review might suffer from a selection bias: due to time and material constrains, we only searched one bibliographic database and a single author selected the studies and extracted the data. Despite snowballing, we might therefore have missed relevant studies. However, including more studies would likely further increase the observed heterogeneity.

With so many differences between the studies, it is difficult to draw firm conclusions about the performance of these algorithms for hospital-level surveillance. Nonetheless, they did appear as useful supplements to traditional surveillance carried out by the infection control teams. In fact, as long as one of these algorithms can detect or confirm outbreaks without generating too many false alarms, it can be considered a useful complementary tool. And because

infection control professionals can more easily investigate an alert than epidemiologists can in the community, a higher rate of false alerts—and therefore a lower PPV—might be acceptable for hospital-level surveillance. But even if high performances are not required, researchers still need a valid and reproducible framework to evaluate and compare these algorithms.

First, researchers need to determine which performance measures they would like to evaluate, as it will have a great impact on the choice of the evaluation approach. Depending on the study design, estimating the four traditional accuracy measures (sensitivity, specificity, positive and negative predictive value) might be difficult. In the context of outbreak detection, however, this may no be an important issue. Indeed, as "non-outbreaks" are by far more frequent than outbreaks, researchers must deal with what are called imbalanced datasets. In these situations, it has been shown that precision-recall plots, based precisely on VPP and sensitivity, give accurate information about classification performance and are more informative than the traditional ROC curves, based on sensitivity and specificity [45]. We also believe that timeliness is an important feature of any outbreak detection system [46,47], and that it should be evaluated along with the traditional detection performance measures. Although some attempts to evaluate timeliness were made in the studies that we included [30,34,36,41], the results are difficult to interpret and future studies should try to address this question more systematically.

The choice of the evaluation approach is also an important aspect of the framework. The evaluation approaches described by Watkins at al. [9] all offer different insights, but the epidemiological approach is preferable for evaluating detection algorithms in real-life settings, because it provides the best face validity. However, the epidemiological approach also has some drawbacks that should be carefully addressed. First, it might suffer from a lack of statistical power, given the relative scarcity of hospital outbreaks in real settings. Sufficiently large surveillance scopes and periods should thus be available to precisely estimate and compare the algorithms' accuracy. A second problem with the epidemiological approach is that it does not provide any obvious reference standard, contrary to the simulation and derived approaches. In the present review, two of the nine studies that followed this approach used a reference standard, but neither of them seemed to us fully satisfactory. Tseng et al. used the results of traditional surveillance [37], which, as shown in some of the included studies, is unfortunately an imperfect reference standard. Hacek et al. used molecular typing [32], but did not apply it to every isolate: it was only considered if an alert was generated, and the final decision was taken by infection control professionals. While this strategy is perfectly understandable from an economic point of view, it introduces a partial verification bias, which leads to an overestimation of the detection performances [29].

The main problem with evaluating detection algorithms is that outbreaks do not have a standard definition [47]. According to recommendations for diagnostic studies by Rutjes et al. [48], one appropriate strategy in such situations is to use expert panels. In order to estimate the validity of the panel's choices, researchers should always report them along with a measure of the inter-rater agreement. Out of the four included studies that used an expert panel, three reported Cohen's kappa coefficients. But, again, differences in how the experts' opinions were collected did not allow direct comparison of the results. The two main approaches were: 1) to ask the experts about the probability that a given alert corresponds to a real outbreak, or 2) to ask the experts what actions should be initiated in response to a given alert. Carnevale et al. [35] used a combination of these two approaches, which complicated the validation process and might partly explain why they measured lower inter-rater agreement. The choice between the two approaches should depend on the expected use of the algorithm. For example, if its purpose is to be used as a practical tool for surveillance, the second approach, which focuses on action, is preferable.

Using experts' knowledge to discern between true and false positives only allows computing VPPs. To estimate the other performance measures, one solution is to use the panel of experts as a real reference standard by asking them to distinguish between "outbreak" and "non-outbreak" periods, as it was done for community outbreak detection in the Bio-ALIRT project [49]. Another solution can be to combine the information brought by various algorithms or data sources. Carnevale et al. [35], for example, gathered the results of different algorithms to estimate the true number of outbreaks and compute sensitivity measures. It is possible however that some outbreaks were missed by all of these sources and that the computed sensitivities were therefore overestimated. In such situations, capture-recapture analyses should be implemented, as proposed in the CDC recommendations for early outbreak detection systems evaluation [47]. Other advanced statistical modeling such as latent class analysis can also combine information from different sources. They are commonly used for diagnostic test evaluation in the absence of a reference standard [48], and it may be interesting to try to use them in the context of outbreak detection.

Another advantage of combining information sources is that it can improve detection performance. For instance, as noted by Carnevale et al. [35], clonal and non-clonal outbreaks have different dynamics and infection control teams may have to use different algorithms to detect each type. This complementarity is well established for control charts: traditional Shewart charts are better for detecting large and sudden deviations from the mean whereas CuSum and exponentially weighted moving averages are better suited for small continuous ones [22].

Several studies also showed that including covariates such as culture site, hospital location and antibiotic resistance can improve the detection performance of the algorithms [23,35,37]. The majority of the studies that we reviewed, however, solely relied on simple time series of cases to trigger alerts. Even though additional sources of data such as electronic medical records and electronic prescriptions might not be readily available for surveillance in all centers, a more thorough investigation of the utility of individual covariates for outbreak detection appears to be another interesting direction for future research.

In addition to a standardized framework, studies on detection algorithms would greatly benefit from a quality assessment tool. We originally wanted to evaluate the quality of the included studies using either the STARD checklist [50] for diagnostic studies or the TRIPOD checklist [51] for prediction model development and validation. Unfortunately, a lot of the items of these tools were not applicable to the context of outbreak detection evaluation. One reason is the variety of the evaluation approaches: the relevant information to be reported is quite different between descriptive, simulation and epidemiological approaches. Another reason is that many items of these checklists address issues about study participants (inclusion, flow, baseline characteristics, adverse events, missing data, etc.), which is not of concern in studies on outbreak detection. Nonetheless, some items of these quality reporting tools address very interesting issues. For example, the TRIPOD checklist differentiates between the development and validation phases for a predictive model. This distinction is important to avoid reporting overly optimistic performances and was only done in seven of the studies that we included. Other examples are the items that relate to statistical power and precision: none of the studies that we included reported a statistical justification of their sample size, and only one of them provided the 95% confidence intervals of their performance measures. Future studies on outbreak detection evaluation should be careful to report these elements.

Undoubtedly, research on the automated detection of hospital outbreak has not yet made the most of the great opportunity offered by modern hospital information systems. More importantly, the evaluation methodology needs to be standardized in order to accurately measure and compare the performances of the detection algorithms. In particular, the different

types of algorithm should be compared in a large study using a valid epidemiological reference standard. With these improvements, we believe that these algorithms can become useful decision-making tools for infection control professionals. They can also help to better understand how outbreaks spread within hospitals, ultimately improving patient safety in healthcare.

Supporting information

S1 Appendix. Pubmed search query.
(DOCX)

S2 Appendix. PRISMA checklist.
(DOC)

Author Contributions

Conceptualization: BL DB.

Data curation: BL.

Formal analysis: BL.

Investigation: BL.

Methodology: BL DB PYB PA.

Software: BL.

Supervision: DL DB.

Visualization: BL.

Writing – original draft: BL DB.

Writing – review & editing: BL DB DL PYB PA.

References

1. Leal J, Laupland KB. Validity of electronic surveillance systems: a systematic review. J Hosp Infect. 2008; 69: 220–229. https://doi.org/10.1016/j.jhin.2008.04.030 PMID: 18550211
2. Murdoch TB, Detsky AS. The Inevitable Application of Big Data to Health Care. JAMA. 2013; 309: 1351–1352. https://doi.org/10.1001/jama.2013.393 PMID: 23549579
3. Leclère B, Lasserre C, Bourigault C, Juvin M-E, Chaillet M-P, Mauduit N, et al. Matching bacteriological and medico-administrative databases is efficient for a computer-enhanced surveillance of surgical site infections: retrospective analysis of 4,400 surgical procedures in a French university hospital. Infect Control Hosp Epidemiol. 2014; 35: 1330–1335. https://doi.org/10.1086/678422 PMID: 25333426
4. Forster AJ, Jennings A, Chow C, Leeder C, van Walraven C. A systematic review to evaluate the accuracy of electronic adverse drug event detection. J Am Med Inform Assoc. 2012; 19: 31–38. https://doi.org/10.1136/amiajnl-2011-000454 PMID: 22155974
5. Kashiouris M, O'Horo JC, Pickering BW, Herasevich V. Diagnostic Performance of Electronic Syndromic Surveillance Systems in Acute Care. Appl Clin Inform. 2013; 4: 212–224. https://doi.org/10.4338/ACI-2012-12-RA-0053 PMID: 23874359
6. Spearing NM, Jensen A, McCall BJ, Neill AS, McCormack JG. Direct costs associated with a nosocomial outbreak of Salmonella infection: an ounce of prevention is worth a pound of cure. Am J Infect Control. 2000; 28: 54–57. PMID: 10679138
7. Mitchell C, Meredith P, Richardson M, Greengross P, Smith GB. Reducing the number and impact of outbreaks of nosocomial viral gastroenteritis: time-series analysis of a multidimensional quality improvement initiative. BMJ Qual Saf. 2015; bmjqs–2015–004134.
8. Buckeridge D, Cadieux G. Surveillance for newly emerging viruses. Persp Med Virol. 2006; 16: 325–43.

9. Watkins RE, Eagleson S, Hall RG, Dailey L, Plant AJ. Approaches to the evaluation of outbreak detection methods. BMC Public Health. 2006; 6: 263. https://doi.org/10.1186/1471-2458-6-263 PMID: 17059615
10. Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004; 140: 189–202. PMID: 14757617
11. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Med. 2009; 6: e1000097. https://doi.org/10.1371/journal.pmed.1000097 PMID: 19621072
12. Childress JA, Childress JD. Statistical test for possible infection outbreaks. Infect Control IC. 1981; 2: 247–249. PMID: 6912215
13. Dessau RB, Steenberg P. Computerized surveillance in clinical microbiology with time series analysis. J Clin Microbiol. 1993; 31: 857–860. PMID: 8463397
14. Mylotte JM. Analysis of infection control surveillance data in a long-term-care facility: use of threshold testing. Infect Control Hosp Epidemiol Off J Soc Hosp Epidemiol Am. 1996; 17: 101–107.
15. Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA. Association rules and data mining in hospital infection control and public health surveillance. J Am Med Inform Assoc JAMIA. 1998; 5: 373–381. PMID: 9670134
16. Arantes A, Carvalho E da S, Medeiros EAS, Farhat CK, Mantese OC. [Use of statistical process control charts in the epidemiological surveillance of nosocomial infections]. Rev Saúde Pública. 2003; 37: 768–774. PMID: 14666307
17. Sagel U, Mikolajczyk RT, Krämer A. Using mandatory data collection on multiresistant bacteria for internal surveillance in a hospital. Methods Inf Med. 2004; 43. Available: http://pub.uni-bielefeld.de/publication/1784036
18. Pentland A, D'Agata EMC, Kumar V. Detecting Clusters of Multidrug-Resistant Gram-Negative Bacteria (MDRGN) using Space-Time Analysis in a Tertiary Care Hospital. Idsa; 2006. https://idsa.confex.com/idsa/2006/webprogram/Paper22552.html
19. Lamma E, Mello P, Nanetti A, Riguzzi F, Storari S, Valastro G. Artificial intelligence techniques for monitoring dangerous infections. IEEE Trans Inf Technol Biomed Publ IEEE Eng Med Biol Soc. 2006; 10: 143–155.
20. Cellarosi G, Lodi S, Sartori C. Detecting outbreaks by time series analysis. Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems, 2002 (CBMS 2002). 2002. pp. 159–164.
21. Menotti J, Porcher R, Ribaud P, Lacroix C, Jolivet V, Hamane S, et al. Monitoring of nosocomial invasive aspergillosis and early evidence of an outbreak using cumulative sum tests (CUSUM). Clin Microbiol Infect. 2010; 16: 1368–1374. https://doi.org/10.1111/j.1469-0691.2009.03150.x PMID: 20041891
22. Gomes IC, Mingoti SA, Di Lorenzo Oliveira C. A novel experience in the use of control charts for the detection of nosocomial infection outbreaks. Clinics. 2011; 66: 1681–1689. https://doi.org/10.1590/S1807-59322011001000004 PMID: 22012038
23. Freeman R, Charlett A, Moore LSP, Davis G, Galletly T, Andrews N, et al. Statistical methods for the prospective detection of outbreaks within the hospital setting: differences in algorithm performance using data available at the national and local levels. European Congress of Clinical Microbiology and Infectious Diseases. Berlin; 2013.
24. Du M, Xing Y, Suo J, Liu B, Jia N, Huo R, et al. Real-time automatic hospital-wide surveillance of nosocomial infections and outbreaks in a large Chinese tertiary hospital. BMC Med Inform Decis Mak. 2014; 14: 9. https://doi.org/10.1186/1472-6947-14-9 PMID: 24475790
25. Faires MC, Pearl DL, Ciccotelli WA, Berke O, Reid-Smith RJ, Weese JS. Detection of Clostridium difficile infection clusters, using the temporal scan statistic, in a community hospital in southern Ontario, Canada, 2006–2011. BMC Infect Dis. 2014; 14: 254. https://doi.org/10.1186/1471-2334-14-254 PMID: 24885351
26. Faires MC, Pearl DL, Ciccotelli WA, Berke O, Reid-Smith RJ, Weese JS. The use of the temporal scan statistic to detect methicillin-resistant Staphylococcus aureus clusters in a community hospital. BMC Infect Dis. 2014; 14: 375. https://doi.org/10.1186/1471-2334-14-375 PMID: 25005247
27. Lefebvre A, Bertrand X, Vanhems P, Lucet J-C, Chavanet P, Astruc K, et al. Detection of Temporal Clusters of Healthcare-Associated Infections or Colonizations with Pseudomonas aeruginosa in Two Hospitals: Comparison of SaTScan and WHONET Software Packages. PLOS ONE. 2015; 10: e0139920. https://doi.org/10.1371/journal.pone.0139920 PMID: 26448036
28. Schifman RB, Palmer RA. Surveillance of nosocomial infections by computer analysis of positive culture rates. J Clin Microbiol. 1985; 21: 493–495. PMID: 3988895

29. Brossette SE, Sprague AP, Jones WT, Moser SA. A data mining system for infection control surveillance. Methods Inf Med. 2000; 39: 303–310. PMID: 11191698

30. Brown SM, Benneyan JC, Theobald DA, Sands K, Hahn MT, Potter-Bynoe GA, et al. Use of Binary Cumulative Sums and Moving Averages in Nosocomial Infection Cluster Detection. Emerg Infect Dis. 2002; 8: 1426–1432. https://doi.org/10.3201/eid0812.010514 PMID: 12498659

31. Ma L, Tsui F-C, Hogan WR, Wagner MM, Ma H. A framework for infection control surveillance using association rules. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2003; 410–414.

32. Hacek DM, Cordell RL, Noskin GA, Peterson LR. Computer-Assisted Surveillance for Detecting Clonal Outbreaks of Nosocomial Infection. J Clin Microbiol. 2004; 42: 1170–1175. https://doi.org/10.1128/JCM.42.3.1170-1175.2004 PMID: 15004070

33. Wright M, Perencevich EN, Novak C, Hebden JN, Standiford HC, Harris AD. Preliminary Assessment of an Automated Surveillance System for Infection Control. Infect Control Hosp Epidemiol. 2004; 25: 325–332. https://doi.org/10.1086/502400 PMID: 15108731

34. Huang SS, Yokoe DS, Stelling J, Placzek H, Kulldorff M, Kleinman K, et al. Automated Detection of Infectious Disease Outbreaks in Hospitals: A Retrospective Cohort Study. PLoS Med. 2010; 7: e1000238. https://doi.org/10.1371/journal.pmed.1000238 PMID: 20186274

35. Carnevale RJ, Talbot TR, Schaffner W, Bloch KC, Daniels TL, Miller RA. Evaluating the utility of syndromic surveillance algorithms for screening to detect potentially clonal hospital infection outbreaks. J Am Med Inform Assoc. 2011; amiajnl–2011–000216.

36. Nishiura H. Early detection of nosocomial outbreaks caused by rare pathogens: a case study employing score prediction interval. Osong Public Health Res Perspect. 2012; 3: 121–127. https://doi.org/10.1016/j.phrp.2012.07.010 PMID: 24159503

37. Tseng Y-J, Wu J-H, Ping X-O, Lin H-C, Chen Y-Y, Shang R-J, et al. A Web-Based Multidrug-Resistant Organisms Surveillance and Outbreak Detection System with Rule-Based Classification and Clustering. J Med Internet Res. 2012; 14: e131. https://doi.org/10.2196/jmir.2056 PMID: 23195868

38. Mellmann A, Friedrich AW, Rosenkötter N, Rothgänger J, Karch H, Reintjes R, et al. Automated DNA Sequence-Based Early Warning System for the Detection of Methicillin-Resistant Staphylococcus aureus Outbreaks. PLoS Med. 2006; 3: e33. https://doi.org/10.1371/journal.pmed.0030033 PMID: 16396609

39. Charvat H, Ayzac L, Girard R, Gardes S, Ecochard R. Detecting related cases of bloodstream infections using time-interval distribution modelling. J Hosp Infect. 2010; 74: 250–257. https://doi.org/10.1016/j.jhin.2009.08.012 PMID: 19914738

40. Kikuchi K, Ohkusa Y, Sugawara T, Taniguchi K, Okabe N. Syndromic Surveillance for Early Detection of Nosocomial Outbreaks. In: Zeng D, Gotham I, Komatsu K, Lynch C, Thurmond M, Madigan D, et al., editors. Intelligence and Security Informatics: Biosurveillance. Springer Berlin Heidelberg; 2007. pp. 202–208. http://link.springer.com/chapter/10.1007/978-3-540-72608-1_20

41. Skipper L. LabGuard: An Automatic Surveillance System for Early Outbreak Detection and Warning of Changes in Antibiotic Resistance Patterns. J Inf Technol Healthc. 2009; 7: 13–22.

42. Wong W-K, Moore A, Cooper G, Wagner M. WSARE: What's Strange About Recent Events? J Urban Health Bull N Y Acad Med. 2003; 80: i66–75.

43. O'Brien SJ, Christie P. Do CuSums have a role in routine communicable disease surveillance? Public Health. 1997; 111: 255–258. PMID: 9242040

44. Kulldorf M, Information Management Services, Inc. SaTScanTM v8.0: Software for the spatial and space-time scan statistics [Internet]. 2009. http://www.satscan.org/

45. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS One. 2015; 10: e0118432. https://doi.org/10.1371/journal.pone.0118432 PMID: 25738806

46. Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW. Algorithms for rapid outbreak detection: a research synthesis. J Biomed Inform. 2005; 38: 99–113. https://doi.org/10.1016/j.jbi.2004.11.007 PMID: 15797000

47. Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V, CDC Working Group. Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the CDC Working Group. MMWR Recomm Rep Morb Mortal Wkly Rep Recomm Rep Cent Dis Control. 2004; 53: 1–11.

48. Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess Winch Engl. 2007; 11: iii, ix–51.

49. Siegrist D, Pavlin J. Bio-ALIRT biosurveillance detection algorithm evaluation. MMWR Suppl. 2004; 53: 152–158. PMID: 15714645

50. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies | Radiology [Internet]. [cited 17 Feb 2017]. http://pubs.rsna.org/doi/abs/10.1148/radiol.2015151516

51. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 2015; 13: 1. https://doi.org/10.1186/s12916-014-0241-z PMID: 25563062