BMJ Open

Evaluations of statistical methods for outlier detection when benchmarking in clinical registries: a systematic review

Jessy Hansen , Susannah Ahern, Arul Earnest

To cite: Hansen J, Ahern S, Earnest A. Evaluations of statistical methods for outlier detection when benchmarking in clinical registries: a systematic review. *BMJ Open* 2023;**13**:e069130. doi:10.1136/bmjopen-2022-069130

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-069130).

Received 13 October 2022
Accepted 26 June 2023

School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia

Correspondence to
Jessy Hansen;
jessy.hansen1@monash.edu

ABSTRACT

Objectives Benchmarking is common in clinical registries to support the improvement of health outcomes by identifying underperforming clinician or health service providers. Despite the rise in clinical registries and interest in publicly reporting benchmarking results, appropriate methods for benchmarking and outlier detection within clinical registries are not well established, and the current application of methods is inconsistent. The aim of this review was to determine the current statistical methods of outlier detection that have been evaluated in the context of clinical registry benchmarking.

Design A systematic search for studies evaluating the performance of methods to detect outliers when benchmarking in clinical registries was conducted in five databases: EMBASE, ProQuest, Scopus, Web of Science and Google Scholar. A modified healthcare modelling evaluation tool was used to assess quality; data extracted from each study were summarised and presented in a narrative synthesis.

Results Nineteen studies evaluating a variety of statistical methods in 20 clinical registries were included. The majority of studies conducted application studies comparing outliers without statistical performance assessment (79%), while only few studies used simulations to conduct more rigorous evaluations (21%). A common comparison was between random effects and fixed effects regression, which provided mixed results. Registry population coverage, provider case volume minimum and missing data handling were all poorly reported.

Conclusions The optimal methods for detecting outliers when benchmarking clinical registry data remains unclear, and the use of different models may provide vastly different results. Further research is needed to address the unresolved methodological considerations and evaluate methods across a range of registry conditions.

PROSPERO registration number CRD42022296520.

BACKGROUND

In recent years, there has been an increase in clinical registry establishment by governments, healthcare administrators and independent bodies worldwide to monitor healthcare quality.[1–3] Clinical registries collect large-scale, prospective health data that is used to support healthcare quality and outcomes improvement. Registries monitor outcomes for numerous and varied patient populations, including specific diseases, events and medical procedures.[3–5] A common purpose of such clinical registries is to monitor the quality of participating healthcare providers, including health systems, sites and individual medical practitioners, to identify underperformers that can be targeted for quality improvement.[2 6–8]

The specific health outcomes monitored by registries vary depending on the patient population and registry purpose, but commonly include mortality, complication and patient-reported outcome measures. Many registries also have clinician agreed clinical quality indicators that can be derived from the collected data,[9 10] which may include process indicators such as time to referral and diagnostic test administration, or clinical and functional measures.

Healthcare provider benchmarking (also called profiling) is often conducted by clinical registries to monitor outcomes. Benchmarking activities involve the comparison of provider performance, at the site or clinician level, and evaluation of variation between such providers. Statistical methods of outlier classification then allow for the identification of underperformers that can be targeted for quality improvement.[3 7 11] An underperformer (or ‘outlier’) within the context of benchmarking health providers refers to a

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ First study to look at evaluations of statistical methods of outlier detection in the context of clinical registries.
⇒ Systematically reviews the literature to identify methodological issues and gaps.
⇒ While a comprehensive search was attempted, some relevant research may have been missed.
⇒ An overall article quality assessment, rather than an assessment of the statistical evaluation quality, was conducted due to the heterogeneity in statistical evaluation techniques used.

provider for which the average outcome deviates from expectation (the benchmark) in the negative direction. Outcomes may be affected by confounding factors outside the control of the health provider, such as the patient demographics of age, socioeconomic status and comorbidity, potentially making the direct comparison of unadjusted outcomes inaccurate. To account for such confounding factors, regression models can be used to compare risk-adjusted estimates. The amount of acceptable deviation from the benchmark is determined by the outlier classification technique. As the amount of variation around an internally derived benchmark is expected be higher for providers with lower case volumes, the categorisation of performance outside expectations is often based on statistical criteria that incorporates sample size, such as CIs and control limits.

A significant body of research has been conducted into benchmarking methodology, including in the context of pay-for-performance programmes and medical insurance data.[12–17] A number of studies have also been conducted to evaluate statistical methods of outlier detection when benchmarking, however, there is little consensus on the most appropriate models and methods, and scarce guidance on which methods should be applied in which setting.[18–27] Clinical registries are distinct from other health data sources to which benchmarking methods have been evaluated due to the unique (non-routine) nature of data collection. Compared with routinely collected data, such as hospital-based systems, registry databases are customised, more comprehensive in collection of clinical data and established with benchmarking as a key purpose. As such, clinical registry data have considerations that must be accounted for when applying methods for benchmarking and outlier detection. These include common limitations such as missing data, overdispersion and low outcome prevalence.[12 22 28–31] Overdispersion occurs when there is large variation in the outcome between providers, which can lead to an over flagging of outliers and make it more difficult to identify true unusual deviation.[28] Outcomes with low prevalence can cause imprecision and uncertainty in results due to low numbers.[22] More specific to registry contexts, the population coverage of the registry, and number and volume of sites, and risk adjustment are important data considerations.[19] The clinical significance, as opposed to statistical significance, and implications of results are also of importance when benchmarking and classifying outliers in clinical registries. Despite these considerations, and the large number of clinical registries conducting benchmarking using a variety of methods, little research has evaluated the best statistical methods of outlier identification within registry settings, as well as the generalisability of methods to different registries.[32] The identification of optimal methods is necessary as different methods have the potential to provide vastly different outcomes. This will allow for consistency and reduce the rate of false positive and negative results. Further, increasing interest in the public reporting of benchmarking results makes it vital that methods are appropriate and robust to ensure accurate information is being communicated to the public, as well as stakeholders.[9 33 34] The potential reputational and employment consequences for medical providers and practitioners from a publicly reported underperformance status is great, adding to the importance of accurate methods.

The purpose of this review is to determine the current statistical methods of outlier detection that have been evaluated in application to clinical registry data (actual or simulated) when benchmarking, assess the benefits and limitations of the methods identified, and determine any gaps in the literature.

METHODS

The systematic review of the literature pertaining to the evaluation of statistical methods of outlier detection in clinical registries described here was conducted to meet the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 Statement guidelines. A PRISMA flow diagram was used to visualise the search strategy and results. The review protocol was prospectively registered on PROSPERO (ID CRD42022296520).

Patient and public involvement

No patient involved.

Search and screening strategy

A search of the literature was conducted using the electronic databases EMBASE, ProQuest, Scopus, Web of Science and Google Scholar. Terms for 'health provider benchmarking', 'clinical registry' and 'outlier detection' were combined with 'AND' to search the databases for relevant articles (full search terms for each database detailed in online supplemental file 1). Searches were restricted to retrieve studies published from 1 January 2010 to capture contemporaneous research, with the final search conducted on 13 April 2023. Database search results were entered into the systematic review platform Covidence (https://www.covidence.org/), which identified and removed duplicates.

One reviewer (JH) screened all titles and abstracts for relevance for progression to the full-text screen and retained relevant reviews for reference list searching. Two independent reviewers (JH and AE) screened the accessible full-text articles against the inclusion criteria to determine eligibility for the review. Author conflicts were resolved through discussion and reasons for exclusion were recorded. One reviewer (JH) searched the reference lists of eligible articles and retained reviews for additional articles for screening.

Eligibility criteria

Articles were eligible for inclusion in the review if they evaluated statistical methods to detect outliers when benchmarking the performance of medical sites, clinicians or devices using actual (or simulated) data from clinical registries. Articles meeting these criteria were included if they were published, peer-reviewed research and had an accessible full-text in English. Study types eligible for inclusion were cross-sectional and cohort studies. Excluded from the review were conference abstracts and editorials, reviews and methods for outlier detection when benchmarking against self or over time. Also excluded were studies using data from administrative datasets, medical records or insurance claims as the different nature of data collection results in more complete patient capture and low missingness, but less targeted clinical data than clinical registries and these datasets were not designed for benchmarking (detailed inclusion and exclusion criteria provided in online supplemental file 2).

Data extraction and quality score tool

Data from eligible articles were extracted into a piloted predefined data extraction form by one reviewer (JH) and checked by another (AE). Data items included citation, registry and outcome information as well as details on the benchmarking and outlier detection methods evaluated in the studies (full list of data items is available in online supplemental file 2).

Two independent reviewers (JH and AE) assessed article quality at a study level using a checklist modified from Harris *et al*,[35] who developed a tool based on Fone *et al*[36] and Jaime Caro *et al*,[37] for the purpose of assessing quality of modelling studies in healthcare. The modified tool assesses 12 criteria on a three-point scale, including the sufficient reporting of data characteristics and methodological detail, and appropriate discussion of assumptions, results and uncertainty (tool criteria, considerations and scores detailed in online supplemental file 2).

Conflicts in data extraction and quality scoring were resolved with discussion. The final consensus data extract and quality score forms were exported from the review platform for summary and synthesis.

Summary measures and evidence synthesis

Data items were summarised in a table to provide an overview of study and registry characteristics included in the review. Details for each study were provided in summary tables (available in online supplemental tables 1 and 2). As no outcome was comparable across all studies, evidence from each study was summarised and presented in a narrative synthesis, grouping studies by the main method comparison or evaluation type. Registry features by methodological category and the characteristics of the most common regression models were also summarised in tables (online supplemental tables 3 and 4), respectively; quality scores were summed and evaluated for each study and criterion (online supplemental table 5).

RESULTS

Study selection

The database search results are presented in figure 1. From the searches, 2093 citations were imported into the review platform, from which 1376 duplicates were removed. Of the 717 titles and abstracts screened, 554 were deemed irrelevant and full texts were accessed for the remaining 163 studies. Of the 163 assessed for eligibility, 144 were

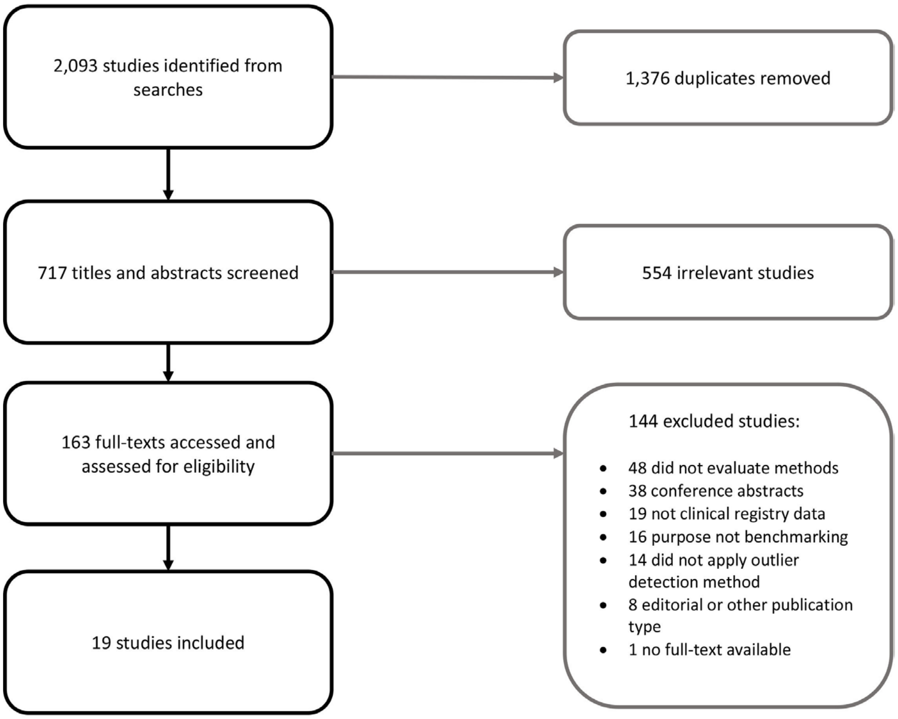
2,093 studies identified from searches
1,376 duplicates removed
717 titles and abstracts screened
554 irrelevant studies
163 full-texts accessed and assessed for eligibility
144 excluded studies:
48 did not evaluate methods
38 conference abstracts
19 not clinical registry data
16 purpose not benchmarking
14 did not apply outlier detection method
8 editorial or other publication type
1 no full-text available
19 studies included

Figure 1 PRISMA flow chart of search results and screening process. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

excluded: 48 did not evaluate methods, 38 were conference abstracts, 19 did not use or simulate clinical registry data, 16 were not for the purpose of benchmarking, 14 did not conduct outlier detection, 8 were an editorial or other publication type and 1 had no full text available. After the full-text screen, 19 studies were deemed eligible for inclusion and had data extracted for the review.

Study and registry characteristics

Summary study (n=19) and registry (n=20) characteristics are presented in table 1 (individual study information provided in additional online supplemental table 1). Of the 19 included studies, 15 conducted application studies,[38–52] while 3 involved applied and simulation studies[53–55] and 1 evaluated a simulation alone.[56] Benchmarking was most commonly evaluated at the site level (90%), although one study assessed both site and clinician benchmarking[52] and another only at the clinician level.[40] Almost all of the included articles used data from only one registry, while one article conducted a simulation study using one registry and an applied analysis to another.[55]

Most included studies used data from registries based in the USA (45%) and Australia and New Zealand (25%). The registries used were commonly established to collect data relating to surgical procedures (50%),[38 40 41 43 47–49 51 54 56] followed by medical events such as stroke or intensive care unit admission (35%),[39 42 44 46 50 53 55] and specific disease registries (15%).[45 52 55] Population coverage was poorly reported with 14 (70%) of the registries having no stated population or site capture; of those that did three were compulsory registries with 100% site coverage,[43 46 51] two estimated over 50% coverage[47 55] and one under 50% coverage.[55] There was a large variation in registry population size (median 39 976 cases; Q1–Q3 9325–149 778) and number of benchmarked groups (median 103 providers; Q1–Q3 55–210).

Mortality was the most common main outcome evaluated (70%), however, there was a range in outcome prevalence between the registries (median 6.7; Q1–Q3 3.8–10.2). All studies assessed outcomes as binary variables, although one study compared the benchmarking of a continuous outcome when analysed as a continuous or dichotomised variable.[44]

Overall study quality as assessed by the modified tool was high, although low scores were common for some criterion (individual study scores presented in online supplemental table 5). Data quality and uncertainty were frequently underaddressed, risk adjustment was often insufficiently described and statistical assumptions were not always explicitly stated. Many studies also failed to appropriately interpret and discuss their results by not considering clinical significance and implications.

Outlier detection methods

Study evaluation and statistical methods information is presented in table 2 (individual study information provided in online supplemental table 2). The included

Table 1 Summary of study and registry characteristics

	N (%)
Study type	
Applied	15 (79.0)
Simulation	1 (5.3)
Applied and simulation	3 (15.8)
Benchmark level	
Site	17 (89.5)
Surgeon/clinician	1 (5.3)
Site and surgeon/clinician	1 (5.3)
Country*	
Australia/New Zealand	5 (25.0)
United States of America	9 (45.0)
Other	6 (30.0)
Registry type*	
Disease	3 (15.0)
Procedure	10 (50.0)
Event	7 (35.0)
Population/site coverage*	
<50%	1 (5.0)
50% to <100%	2 (10.0)
100% (compulsory)	3 (15.0)
Not stated	14 (70.0)
Population size in analysis*	
<10 000	5 (25.0)
10 000 to <25 000	2 (10.0)
25 000 to <50 000	2 (10.0)
50 000 to <1 00 000	2 (10.0)
100 000 to <5 00 000	4 (20.0)
≥500 000	2 (10.0)
Not stated	3 (15.0)
Median (IQR)	39 976 (9325–149 778)
Range	1815–1 984 998
No of benchmarked groups*	
<50	4 (20.0)
50 to <100	6 (30.0)
100 to <500	6 (30.0)
≥500	4 (20.0)
Median (IQR)	102.5 (55.25–210.25)
Range	8–1292
Main outcome*	
Mortality	14 (70.0)
Other	6 (30.0)
Prevalence (main outcome)*	
<2.5%	4 (20.0)
2.5% to <5%	2 (10.0)
5% to <10%	5 (25.0)

Continued

Table 1 Continued

	N (%)
10% to <20%	4 (20.0)
≥20	2 (10.0)
NA/not stated	3 (15.0)
Median (IQR)	6.7 (3.8–10.2)
Range	1.4–23.2
Outcome data type†	
Binary	19 (100)
Continuous	1 (5.3)

*One study has two registries, n=20.
†Multiple possible, row per cent.
IQR, interquartile range; NA, not applicable.

studies evaluated a wide range of statistical models and outlier classification methods. A common evaluation was comparing random effects (RE) to fixed effects (FE) models (26%),[41,45,52,53,55] or the comparison of hierarchical to ordinary models (11%).[38,46] Assessing multiplicity adjustment,[42,43,50] the evaluation of one method across varying data scenarios,[44,49,56] and comparing results from Bayesian and frequentist frameworks were each the focus of some studies[48,54] (16%, 16% and 11%, respectively).

There was variation in the models recommended or used, if evaluating a different methodological aspect, by each study. Hierarchical and RE logistic regression were the most commonly used (47%),[38,40–43,45–47,50,57] applied in both frequentist and Bayesian frameworks, followed by FE and ordinary logistic regression (32%).[45,49,51–53,55] Cox regression was used in three studies: one each applying ordinary,[48] random effects[56] and Bayesian regression.[54] One study applied a hierarchical generalised additive model[39] while another used no model,[44] instead applying outlier detection methods to the raw outcome rates. Of the seven studies evaluating hierarchical or RE regression compared with ordinary or FE, three studies recommended or used hierarchical models for their final results,[38,41,46] three studies found ordinary or FE regression to perform better[52,53,55] and one study concluded both models to be appropriate if adequately risk adjusted[45] (the strengths and limitations for these models as identified by the studies are presented in online supplemental table 4).

While all studies used an internal benchmark, the outlier detection method used or recommended by the studies varied. The most common outlier classification method was to use 95% CIs calculated for individual providers (42%), either for raw or standardised outcome rates and ratios,[38,41,45,49,51,53] or ORs.[40,46] Funnel plot limits based on data parameters of an internally derived benchmark and provider volumes were also commonly used to detect outliers (37%), with limits determined using SD and 95% CIs[39,44,47,52] or controlling the false discovery rate to adjust for multiplicity.[42,43,50] Two studies applied

Table 2 Summary of statistical methods

	N (%)
Main comparison	
Random versus fixed effects	5 (26.3)
Multiple comparisons adjustment	3 (15.8)
One method across data frames	3 (15.8)
Bayesian versus frequentist	2 (10.5)
Hierarchical versus ordinary	2 (10.5)
Other	4 (21.1)
Method recommended/used†	
Hierarchical logistic regression	5 (26.3)
Bayesian hierarchical logistic regression	2 (10.5)
Random effects logistic regression	2 (10.5)
Fixed effects logistic regression	4 (21.1)
Ordinary logistic regression	2 (10.5)
Cox regression	1 (5.3)
Random effects Cox regression	1 (5.3)
Bayesian Cox regression	1 (5.3)
Hierarchical generalised additive model	1 (5.3)
Nil (raw rates)	1 (5.3)
Outlier classification method recommended/used	
Estimate 95% CIs	6 (31.6)
OR 95% CIs	2 (10.5)
2/3 SD or 95% CI funnel plot limits	4 (21.1)
5% false discovery rate funnel plot limits	3 (15.8)
Clinical and probabilistic thresholds	2 (10.5)
Mixed criteria	2 (10.5)
Risk-adjustment model build*	
Model fit/internal validation	5 (25.0)
Stepwise	4 (20.0)
Expert opinion (a priori)	1 (5.0)
None	1 (5.0)
Not stated	7 (35.0)
Simulation (assume sufficient)	2 (10.0)
Missing data handling*	
Complete case	8 (40.0)
Multiple imputation	1 (5.0)
Mixed methods	2 (10.0)
Not stated	7 (35.0)
NA (simulation)	2 (10.0)
Group volume minimum*	
No minimum	6 (30.0)
10	4 (20.0)
50	2 (10.0)
150	2 (10.0)
Not stated	6 (30.0)
Visualisation†	

Continued

Table 2 Continued

	N (%)
Funnel plot	7 (35.0)
Caterpillar plot	6 (30.0)
Other	3 (15.0)
None	4 (20.0)

*One study has two registries, n=20.
†Multiple possible, row per cent.
CI, confidence interval; NA, not applicable; OR, odds ratio; SD, standard deviation.

methods using a Bayesian framework that incorporated clinical and probabilistic thresholds to determine outliers,[54 55] while another two used mixed criteria based on observed and expected values.[48 56]

Risk adjustment was conducted in all but one study[44] to obtain standardised estimates for benchmarking. A statistically based model development technique was most common (45%), with studies using model fit and internal validation[39 42 46 51 52] or stepwise variable inclusion[38 41 49 50] to build the risk adjustment model. Several studies did not describe the risk adjustment build process despite applying a risk adjusted model (35%).[40 43 45 48 53–55]

Missing data were handled differently across the studies, but was poorly reported, with a large number of articles not addressing missing data (35%).[38 40 45 47 48 52 54] Of those with sufficient descriptions, complete case was the most common method (40%),[39 41–44 49 50 55] while some studies incorporated multiple imputation techniques (15%).[46 51 53] Another poorly reported methodological consideration was provider volume minimum, with almost a third of studies providing insufficient information.[39 44–46 51 56] For the remaining studies, there was variation in the minimum case volume for providers to be included in the analysis. No minimum was the most common (30%),[38 40 52–55] followed by 10 (20%),[43 47 48 55] 50[41 49] and 150 case minimums[40 50] (10%, each).

Different graph types were used to visualise benchmarking and outlier detection, with funnel (35%)[39 42–44 47 50 52] and caterpillar plots (30%)[40 45 46 50 51 53] being the most common, although some studies presented no graphical display of the results (20%).[38 41 48 54]

Registry characteristics such as population size and number of providers were similar when stratified by regression model and outlier classification method categories. Random effects regression was only evaluated in an applied framework and not assessed for any registries with a very low (<2.5%) outcome prevalence. A geographical difference was observed for outlier classification method, with more studies using registries from the USA using CIs around provider estimates to flag outliers, and registries from Australia and other nations applying control limits.

DISCUSSION

The review of studies comparing statistical methods for outlier detection when benchmarking clinical registry data presented here found a small body of literature, leading to inconclusive results regarding the optimal methods. Most of the identified studies evaluated methods by comparing the outcome estimates and number of outliers detected without robust assessment of the accuracy of the outliers and appropriateness of the method. Few of the studies conducted more rigorous method performance evaluations through the use of simulated data sets with a known set of 'true' outliers. A wide range of benchmark models and outlier classification methods were applied and evaluated in the studies. A common comparison among the articles assessing hierarchical or RE models against ordinary or FE had mixed results. From the research identified in this systematic review, there appears to be little consensus as to the best methods for outlier detection in clinical registries, including the validity, accuracy and calibration of outcomes using these methods. In addition, the review identified several important data and analysis considerations that are underaddressed in the current literature. Simulation studies addressing these issues and considerations are needed to determine the effectiveness of methods in various settings, including across a range of outcome prevalence sample size as observed in our review, to find the most appropriate methods for outlier detection in clinical registries.

Validity of the measurement (regression model)

While almost all of the included studies applied a regression model to risk adjust outcome estimates, the validity of the estimates for each model was not clear based on the included studies due to the heterogeneity of the evaluated models and the lack of robust simulation studies allowing for a comparison to a known 'true value'.

For many studies not comparing statistical models, such as those that looked at different outlier classifications, hierarchical logistic regression was chosen for analysis. Despite its frequent use, there was little evidence cited to support the choice of model beyond theoretically higher accuracy, and it was preferred in less than half of the studies comparing it to ordinary or FE regression. For the studies recommending hierarchical regression, many failed to discuss the limitations or address the statistical assumptions of the model, and none evaluated the method using a simulation study. There is much discussion about the use and appropriateness of hierarchical and RE logistic regression when applied to detecting outliers when benchmarking.[13 15 26 58 59] Allowing for provider level effects causes estimate shrinkage and a reduction in variation, and can be used as a method to handle overdispersion.[25] Some have argued that hierarchical modelling 'adjusts for reliability' and allows for increased variation for lower volume sites, potentially reducing the risk of false flagging underperformers.[60] For the studies in this review, hierarchical regression generally did result in a reduction in provider variation and fewer outliers detected. Other studies, however, have found the attenuation from RE regression to bias estimates towards the null,[15] and cause poorer outlier detection performance by increasing the false negative rate,[58 61] especially for low volume sites. The allowance for provider level effects in hierarchical regression may be appropriate when producing risk prediction models or patient stratification for use in a clinical setting. However, the benefit of a large reduction in variation when the purpose is to detect outliers, often defined based on expected variance criteria, is not established. In addition, the use of hierarchical models with funnel plots to detect outliers may be inappropriate as it 'double-dips' on allowance for provider volume, which may result in further false negatives. The use of such models may be appealing as they provide less contentious results that are more acceptable to clinicians. Ordinary or FE models were preferred by almost half of the studies comparing it to RE regression, with better computation time and modest statistical advantages such as higher power given as justification. The lack of shrinkage has been argued by some to be a limitation of FE modelling, however, as it produces less stable estimates and is more prone to overdispersion.[45 46] Other studies have found the lack of distributional assumption and absence of shrinkage to result in less estimate bias and better sensitivity.[15 58 62]

Some studies included in the review evaluated outliers detected from methods in a Bayesian framework compared with frequentist methods. Many of these were comparing to traditional outlier classification techniques based on mixed criteria for the expected and observed outcome estimates, and some found Bayesian methods to perform better. However, many limitations to using Bayesian models for outlier detection were identified in the review, and have been discussed in the broader literature.[63–65] The high computation time and difficulty in conducting Bayesian analyses, including the challenge of selecting appropriate prior distributions, are barriers to implementation in clinical registries. Further, the results can be more difficult to interpret and communicate to clinicians, stakeholders and the public.

Accuracy of the measurement (dispersion estimation)

Beyond the references to 'reliability adjustment' and estimate shrinkage through the application of hierarchical regression models, no mention of the accuracy of the outcome measurements were made by the included studies. The methods of estimating dispersion were not directly addressed and most studies did not report a SD of the outcome at the health provider level.

Calibration of the measurement (outlier classification method)

While many studies evaluated outlier detection across different models using the same outlier definition, very few studies directly evaluated the validity of the calibration methods for comparing health provider estimates when benchmarking. Of those that did assess the performance of different outlier categorisation techniques, most looked at adjusting for multiplicity when constructing limits on funnel plots. Due to the high number of comparisons made when benchmarking, multiplicity may affect the significance of results and artificially increase the number of outliers identified.[66] Controlling for multiple comparisons reduced the number of outliers detected for the studies in this review, and these results were preferred by the studies as they accounted for the problem of multiplicity caused by comparing numerous providers against a benchmark, however, none of these evaluations were conducted in a simulation study allowing for generalisability across registry settings and outcomes.

Gaps in the included literature

Despite the population capture of registries having implications on the certainty of results, population coverage, either at the site or patient level, was not reported for the majority of studies in this review. The population coverage of registries is often changing which can bias the estimated 'population average', often used as a benchmark. An over or underestimated benchmark may cause errors in outlier detection by classifying outliers based on the sample captured by the registry, which may not reflect their performance status compared with the whole population, and limits result generalisability. None of the studies included in the review, even those that reported population coverage, evaluated or addressed this source of benchmark uncertainty. Studies have found outlier detection without accounting for benchmark uncertainty to have higher false positive rates, and recommend the use of tolerance or uncertainty intervals.[19 67] As outliers are most frequently detected using methods comparing provider performance to an internally derived benchmark, it is vital that potential uncertainty from population coverage is sufficiently addressed, either in the methods to detect outliers or in the interpretation of results.

The use of a provider case volume minimum was another poorly reported factor that can affect the results of benchmarking. Among the studies that reported their case volume minimum, there was variation in those applied, ranging from no minimum to 150. The inclusion of low volume sites is an important analysis consideration, however, only one of the included studies evaluated results from different minimums, although it found little difference. While low volume sites can be expected to have higher variation, and may be flagged due to random chance rather than true underperformance, simply excluding them will mean that any poor performance by sites with low case volume cannot be detected at all. Their exclusion may also have implications on internally derived benchmarks by changing the mean or median outcome. The appropriateness of a case volume minimum for inclusion in benchmarking and outlier detection analyses needs to be further researched.

Further important considerations were also underaddressed in the included literature. Clinical significance was rarely considered when classifying outlier and interpreting the results, despite its importance to ensuring clinically relevant outliers are detected. Most studies evaluated outcomes with low prevalence, however, many clinical quality indicators have higher prevalence and method performance can vary for different outcome prevalence.[58]

Strengths and limitations

This review is the first, to our knowledge, to identify the gaps in the literature regarding the optimal model choice for detecting outliers when benchmarking in clinical registries. Other studies have reviewed statistical methodology more broadly and in other healthcare contexts, or focused only on the application of methods, while this review assessed evaluation studies. Though a comprehensive search was attempted, the heterogeneity of terminology for benchmarking and outliers may have resulted in some studies being missed by the search strategy. The variety in statistical evaluation method made finding an appropriate quality score tool difficult, and as such a more general article quality assessment was conducted.

Conclusions

Clinical registries are important resources for monitoring healthcare quality through benchmarking outcomes, including process and outcome quality indicators. The rise in clinical registry data and interest in public reporting mean that provider outlier detection is high stakes and robust methodology is vital. This review of evaluations of statistical methods for outlier detection within clinical registries found many unresolved considerations, including model choice, outlier definition, benchmark uncertainty and case volume minimum. Most of the studies included in the review conducted simple comparisons, looking at the number of outliers categorised by each method without rigorous statistical evaluation. This review highlights that much uncertainty remains as to the most appropriate benchmarking and outlier detection method for use in clinical registries, and many important analytical considerations of registries have not been sufficiently addressed. Given this, future research should look to use simulations studies to robustly evaluate statistical method performance under a range of registry conditions and data limitations, including varying prevalence, sample size and dispersion, to ensure valid, generalisable and reproducible results. Such research would provide a framework for analysis and reporting of outliers specific registry settings, as there may not be a global optimal model.

Contributors JH and AE designed the study. JH designed the search strategy, conducted the systematic search and screened titles and abstracts. JH and AE screened and reviewed full texts, and quality scored articles. Data extraction was performed by JH and checked by AE. JH synthesised the evidence and drafted the manuscript. AE and SA critically reviewed and edited the manuscript. All authors read and approved the final manuscript and AE is the guarantor of the work.

Funding JH is supported by an Australian Government Research Training Programme (RTP) Stipend and RTP Fee-Offset Scholarship through Monash University. This research received no other specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Ethics approval was not required as the research is based exclusively on published literature.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplemental information.

ORCID iDs

Jessy Hansen http://orcid.org/0000-0001-5583-8773
Arul Earnest http://orcid.org/0000-0003-2693-5034

REFERENCES

1 Lecky F, Woodford M, Edwards A, *et al*. Trauma scoring systems and databases. *Br J Anaesth* 2014;113:286–94.
2 McNeil JJ, Evans SM, Johnson NP, *et al*. Clinical-Quality registries: their role in quality improvement. *Med J Aust* 2010;192:244–5.
3 Hoque DME, Kumari V, Hoque M, *et al*. Impact of clinical registries on quality of patient care and clinical outcomes: A systematic review. *PLoS One* 2017;12:e0183667.
4 Evans SM, Bohensky M, Cameron PA, *et al*. A survey of Australian clinical registries: can quality of care be measured *Intern Med J* 2011;41:42–8.
5 Stey AM, Russell MM, Ko CY, *et al*. Clinical registries and quality measurement in surgery: a systematic review. *Surgery* 2015;157:381–95.
6 Blumenthal S. n.d. The use of clinical registries in the United States: A landscape survey. *eGEMs*;5:26.
7 Wilcox N, McNeil JJ. Clinical quality registries have the potential to drive improvements in the appropriateness of care. *Med J Aust* 2016;205:S27–9.
8 Brown WA, Ahern S, MacCormick AD, *et al*. Clinical quality registries: urgent reform is required to enable best practice and best care. *ANZ J Surg* 2022;92:23–6. 10.1111/ans.17438 Available: https://onlinelibrary.wiley.com/toc/14452197/92/1-2
9 Ahern S, Hopper I, Evans SM. Clinical quality registries for clinician-level reporting: strengths and limitations. *Med J Aust* 2017;206:427–9.
10 Evans SM, Scott IA, Johnson NP, *et al*. Development of clinical-quality registries in Australia: the way forward. *Med J Aust* 2011;194:360–3.
11 Blackmore AR, Leonard J, Madayag R, *et al*. Using the trauma quality improvement program Metrics data to enhance clinical practice. *J Trauma Nurs* 2019;26:121–7.
12 Spiegelhalter D, Sherlaw-Johnson C, Bardsley M, *et al*. Statistical methods for Healthcare regulation: rating, screening and surveillance: statistical methods for Healthcare regulation. *J R Stat Soc Ser A Stat Soc* 2012;175:1–47.
13 Eijkenaar F, van Vliet RCJA. Performance profiling in primary care: does the choice of statistical model matter *Med Decis Making* 2014;34:192–205.
14 Ieva F, Paganoni AM. Detecting and Visualizing Outliers in provider profiling via funnel plots and mixed effect models. *Health Care Manag Sci* 2015;18:166–72.
15 Kalbfleisch JD, Wolfe RA. On monitoring outcomes of medical providers. *Stat Biosci* 2013;5:286–302.

16 Normand S-LT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc* 1997;92:803–14.

17 Guglielmi A, Ieva F, Paganoni AM, *et al*. Semiparametric Bayesian models for clustering and classification in the presence of unbalanced in-hospital survival. *J Royal Stat Soc Series C* 2014;63:25–46.

18 Racz MJ, Sedransk J. Inference for identifying outlying health care providers. *J Statist Plan Infer* 2015;160:51–9.

19 Manktelow BN, Seaton SE, Evans TA. Funnel plot control limits to identify poorly performing Healthcare providers when there is uncertainty in the value of the benchmark. *Stat Methods Med Res* 2016;25:2670–84.

20 Longford NT. Decision theory for comparing institutions. *Stat Med* 2018;37:457–72.

21 Austin PC, Naylor CD, Tu JV. A comparison of a Bayesian vs. a Frequentist method for profiling hospital performance. *J Eval Clin Pract* 2001;7:35–45.

22 Seaton SE, Barker L, Lingsma HF, *et al*. What is the probability of detecting poorly performing hospitals using funnel plots *BMJ Qual Saf* 2013;22:870–6.

23 Austin PC. A comparison of Bayesian methods for profiling hospital performance. *Med Decis Making* 2002;22:163–72.

24 Paddock SM, Louis TA. Percentile-based empirical distribution function estimates for performance evaluation of Healthcare providers. *J Royal Statist Soc Ser C* 2011;60:575–89.

25 Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Statist Med* 2005;24:1185–202. 10.1002/sim.1970 Available: http://doi.wiley.com/10.1002/sim.v24:8

26 Ohlssen DI, Sharples LD, Spiegelhalter DJ. A Hierarchical Modelling framework for identifying unusual performance in health care providers. *J Royal Stat Soc Series A* 2007;170:865–90.

27 Verburg IWM, Holman R, Peek N, *et al*. Guidelines on constructing funnel plots for quality indicators: A case study on mortality in intensive care unit patients. *Stat Methods Med Res* 2018;27:3350–66.

28 Spiegelhalter DJ. Handling over-dispersion of performance indicators. *Qual Saf Health Care* 2005;14:347–51.

29 Walker K, Neuburger J, Groene O, *et al*. Public reporting of surgeon outcomes: low numbers of procedures lead to false complacency. *Lancet* 2013;382:1674–7.

30 Psoter KJ, Rosenfeld M. Opportunities and pitfalls of Registry data for clinical research. *Paediatr Respir Rev* 2013;14:141–5.

31 Thompson MP, Luo Z, Gardiner J, *et al*. Impact of missing stroke severity data on the accuracy of hospital ischemic stroke mortality profiling: A simulation study. *Circ Cardiovasc Qual Outcomes* 2018;11:e004951.

32 Chung G, Etter K, Yoo A. Medical device active surveillance of spontaneous reports: A literature review of signal detection methods. *Pharmacoepidemiol Drug Saf* 2020;29:369–79.

33 Thompson MR, Tekkis PP, Stamatakis J, *et al*. The National bowel cancer audit: the risks and benefits of moving to open reporting of clinical outcomes. *Colorectal Dis* 2010;12:783–91.

34 Behrendt K, Groene O. Mechanisms and effects of public reporting of surgeon outcomes: A systematic review of the literature. *Health Policy* 2016;120:1151–61.

35 Harris RC, Sumner T, Knight GM, *et al*. Systematic review of mathematical models exploring the Epidemiological impact of future TB vaccines. *Hum Vaccin Immunother* 2016;12:2813–32.

36 Fone D, Hollinghurst S, Temple M, *et al*. Systematic review of the use and value of computer simulation Modelling in population health and health care delivery. *J Public Health Med* 2003;25:325–35.

37 Jaime Caro J, Eddy DM, Kan H, *et al*. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC good practice task force report. *Value in Health* 2014;17:174–82.

38 Dimick JB, Ghaferi AA, Osborne NH, *et al*. Reliability adjustment for reporting hospital outcomes with surgery. *Health Serv Res* 2012;255:703–7.

39 Endo H, Uchino S, Hashimoto S, *et al*. Development and validation of the predictive risk of death model for adult patients admitted to intensive care units in Japan: an approach to improve the accuracy of Healthcare quality measures. *J Intensive Care* 2021;9:18.

40 Hall BL, Huffman KM, Hamilton BH, *et al*. Profiling individual surgeon performance using information from a high-quality clinical Registry: opportunities and limitations. *J Am Coll Surg* 2015;221:901–13.

41 Hess CN, Rao SV, McCoy LA, *et al*. Identification of hospital Outliers in bleeding complications after percutaneous coronary intervention. *Circ Cardiovasc Qual Outcomes* 2015;8:15–22.

42 Kasza J, Moran JL, Solomon PJ, *et al*. Evaluating the performance of Australian and New Zealand intensive care units in 2009 and 2010. *Stat Med* 2013;32:3720–36.

43 Kasza J, Polkinghorne KR, Wolfe R, *et al*. Comparing dialysis centre mortality outcomes across Australia and New Zealand: identifying unusually performing centres 2008-2013. *BMC Health Serv Res* 2018;18:1007.

44 Kuhrij L, van Zwet E, van den Berg-Vos R, *et al*. Enhancing feedback on performance measures: the difference in Outlier detection using a binary versus continuous outcome funnel plot and implications for quality improvement. *BMJ Qual Saf* 2021;30:38–45.

45 MacKenzie TA, Grunkemeier GL, Grunwald GK, *et al*. A primer on using shrinkage to compare in-hospital mortality between centers. *Ann Thorac Surg* 2015;99:757–61.

46 Moore L, Hanley JA, Turgeon AF, *et al*. Evaluating the performance of trauma centers: Hierarchical modeling should be used. *J Trauma* 2010;69:1132–7.

47 Penninckx F, Beirens K, Fieuws S, *et al*. Risk adjusted Benchmarking of clinical anastomotic leakage rate after total Mesorectal Excision in the context of an improvement project. *Colorectal Dis* 2012;14:e413–21.

48 Schold JD, Miller CM, Henry ML, *et al*. Evaluation of flagging criteria of United States kidney transplant center performance: how to best define Outliers *Transplantation* 2017;101:1373–80.

49 Sherwood MW, Brennan JM, Ho KK, *et al*. The impact of extreme-risk cases on hospitals' risk-adjusted percutaneous coronary intervention mortality ratings. *JACC: Cardiovascular Interventions* 2015;8:10–6.

50 Solomon PJ, Kasza J, Moran JL, *et al*. Identifying unusual performance in Australian and New Zealand intensive care units from 2000 to 2010. *BMC Med Res Methodol* 2014;14:53.

51 Spertus JV, T Normand S-L, Wolf R, *et al*. Assessing hospital performance after percutaneous coronary intervention using big data. *Circ Cardiovasc Qual Outcomes* 2016;9:659–69.

52 Teloken PE, Heriot AG, Hunter A, *et al*. Analysis of mortality in colorectal surgery in the bi-national colorectal cancer audit. *ANZ J Surg* 2016;86:951–2.

53 Moran JL, Solomon PJ, ANZICS Centre for Outcome and Resource Evaluation (CORE) of Australian and New Zealand Intensive Care Society (ANZICS). Fixed effects Modelling for provider mortality outcomes: analysis of the Australia and New Zealand intensive care society (ANZICS) adult patient data-base. *PLoS ONE* 2014;9:e102297.

54 Salkowski N, Snyder JJ, Zaun DA, *et al*. A scientific Registry of transplant recipients Bayesian method for identifying underperforming transplant programs. *American Journal of Transplantation* 2014;14:1310–7.

55 Varewyck M, Goetghebeur E, Eriksson M, *et al*. On shrinkage and model extrapolation in the evaluation of clinical center performance. *Biostatistics* 2014;15:651–64.

56 Massie AB, Segev DL. Rates of false flagging due to statistical Artifact in CMS evaluations of transplant programs: results of a stochastic simulation: rates of false flagging in CMS evaluations. *Am J Transplant* 2013;13:2044–51.

57 Hess BJ, Weng W, Lynn LA, *et al*. Setting a fair performance standard for physicians' quality of patient care. *J Gen Intern Med* 2011;26:467–73.

58 Austin PC, Alter DA, Tu JV. The use of Fixed- and random-effects models for classifying hospitals as mortality Outliers: a Monte Carlo assessment. *Med Decis Making* 2003;23:526–39.

59 Jones HE, Spiegelhalter DJ. The identification of "unusual" health-care providers from a Hierarchical model. *The American Statistician* 2011;65:154–63.

60 Dimick JB, Staiger DO, Birkmeyer JD. Ranking hospitals on surgical mortality: the importance of reliability adjustment: ranking hospitals on surgical mortality. *Health Serv Res* 2010;45:1614–29.

61 Racz MJ, Sedransk J. Bayesian and Frequentist methods for provider profiling using risk-adjusted assessments of medical outcomes. *J Am Stat Assoc* 2012;105:48–58.

62 Kipnis P, Escobar GJ, Draper D. Effect of choice of estimation method on inter-hospital mortality rate comparisons. *Med Care* 2010;48:458–65.

63 Wong AYL, Warren S, Kawchuk GN. A new statistical trend in clinical research - Bayesian Statistics. *Physical Therapy Reviews* 2010;15:372–81.

64 Hackenberger BK. Bayes or not Bayes, is this the question *Croat Med J* 2019;60:50–2.

65 van de Schoot R, Kaplan D, Denissen J, *et al*. A gentle introduction to Bayesian analysis: applications to developmental research. *Child Dev* 2014;85:842–60.

66 Jones HE, Ohlssen DI, Spiegelhalter DJ. Use of the false discovery rate when comparing multiple health care providers. *J Clin Epidemiol* 2008;61:232–40.

67 Paddock SM. Statistical benchmarks for health care provider performance assessment: A comparison of Standard approaches to a Hierarchical Bayesian Histogram-based method. *Health Serv Res* 2014;49:1056–73.