Title: **Dynamic methods for ongoing assessment of site-level risk in risk-based monitoring of clinical trials: a scoping review**

Running Head: Dynamic methods for site-level risk assessment

Version + Date: Version 1.0, 29th February 2020

Word count: Abstract: 409

Text (excluding abstract, references, appendix, tables and figures): 3984

Tables/Figures: 4 tables, 2 figures

Supplement: 3 files

Authors:	**Name**	**Affiliation**
	William J Cragg ORCID: 0000-0002-1274-8521	1: MRC Clinical Trials Unit at UCL, London, UK; 2: Clinical Trials Research Unit, Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, UK
	Caroline Hurley ORCID: 0000-0002-3793-3408	Health Research Board-Trials Methodology Research Network (HRB-TMRN), National University of Ireland, Galway, Ireland
	Victoria Yorke-Edwards ORCID: 0000-0002-6465-7330	MRC Clinical Trials Unit at UCL, London, UK
	Sally P Stenning ORCID: 0000-0001-7009-5488	MRC Clinical Trials Unit at UCL, London, UK

Correspondence: William Cragg

w.cragg@leeds.ac.uk

Tel: 0113 343 8398

Clinical Trials Research Unit

Leeds Institute of Clinical Trials Research

University of Leeds

Leeds

LS2 9JT,

United Kingdom

Research support: Medical Research Council (MC_UU_12023/24)

The TEMPER study was funded by a grant from Cancer Research UK (C1495/A13305)

Summary of figures and tables

Table 1: general characteristics of included studies

Table 2: full listing of all included reports

Table 3: types of assessments and evidence presented by reports that included some assessments of their methods' effectiveness

Table 4: best reported information on methods' classification ability, where available or deductible

Figure 1: PRISMA flow diagram

Figure 2: publications by year and type

Summary of supplementary information

1. Completed PRISMA-ScR Checklist
2. Search strategy from Medline
3. Annotated list of data collection variables

1

Abstract

Background/Aims

It is increasingly recognised that reliance on frequent site visits for monitoring clinical trials is inefficient. Regulators and trialists have in recent years encouraged more risk-based monitoring. Risk assessment should take place before a trial begins in order to define the overarching monitoring strategy. It can also be done on an ongoing basis, in order to target sites for monitoring activity. Various methods have been proposed for such prioritisation, often using terms like 'central statistical monitoring', 'triggered monitoring' or, as in ICH Good Clinical Practice guidance, ‘targeted on-site monitoring’. We conducted a scoping review to identify such methods, to establish if any published methods were supported by adequate evidence to allow wider implementation, and to point the way to future developments in this field of research.

Methods

We used 7 publication databases, 2 sets of methodological conference abstracts and an internet search engine to look for methods for using centrally held trial data to assess site conduct during a trial. We included only reports in English, and excluded reports published before 1996 and reports not directly relevant to our research question. We used reference and citation searches to find additional relevant reports. We extracted data using a pre-defined template. We contacted authors to request additional information about included reports and to check whether reports might be eligible.

Results

We included 30 reports in our final dataset, of which 21 were peer-reviewed publications. 20 reports described central statistical monitoring methods (of which 7 focussed on

detection of fraud or misconduct) and 9 described triggered monitoring methods. 21 reports included some assessment of their methods' effectiveness. Most commonly this involved exploring the methods' characteristics using real trial data with no known integrity issues. Of the 21 with some effectiveness assessment, most presented limited or no information about whether or not concerns identified through central monitoring constituted meaningful problems. Some reports commented on cost savings from reduced on-site monitoring, but none gave detailed costings for the development and maintenance of central monitoring methods themselves.

Conclusions

Our review identified various proposed methods, some of which could be combined within the same trial. The apparent emphasis on fraud detection may not be proportionate in all trial settings. Although some methods have self-justifying benefits for data cleaning activity, many have limitations that may currently prevent their routine use for targeting trial monitoring activity. The implementation costs, or uncertainty about these, may also be a barrier. We make recommendations for how the evidence-base supporting these methods could be improved.

Keywords: trial monitoring, risk-based monitoring, triggered monitoring, central statistical monitoring, Good Clinical Practice, research misconduct, data fabrication

Introduction

Monitoring, a major component of assuring the quality of clinical trials, has traditionally relied on frequent on-site monitoring visits,[1] particularly to facilitate sometimes extensive source data verification (SDV).[2] However, it is increasingly recognised that this model may be inefficient and unnecessary in many cases,[3,4] with trialists questioning the value of 100% SDV.[5–7] In recent years, regulators[8–10] and trialists[1,11] have proposed a risk-based approach to monitoring, whereby monitoring methods, including the frequency and nature of on-site visits, vary across trials depending on the risks specific to each one.

Risk-based monitoring methods can be applied at different stages of a trial. Pre-trial risk assessments can help define the overarching strategies appropriate to the trial's risks. In some models,[12,13] this is predominantly a one-off assessment during trial setup. However, it is also possible to modify the monitoring strategy, or build in flexibility and responsiveness, based on changing or emerging risks during the course of the trial.[14]

Risk-based monitoring is often associated with fewer on-site visits than 'traditional' monitoring.[15] Although effective central monitoring methods alone could, in some respects, provide adequate trial monitoring in place of visits, on-site visits offer particular benefits over central monitoring, for example in any activities that require access to site-held source data (such as patients' medical notes). On-site visits may also be necessary, for example, to investigate potential fraudulent activity. In a risk-based monitoring framework, visits to sites may not be routine, but can be based on assessed risk; we therefore need methods to assess site-level risk on an ongoing basis. We can interpret these methods as assessing the risk of *not going to site now.* If the risk seems too high, a visit – or some other corrective action – is triggered. Methods of this kind have been referred to using various terms,

including 'triggered monitoring'[14] or, as in ICH GCP guidance, 'targeted monitoring',[16] and may employ data-driven approaches from methods known collectively as 'central statistical monitoring',[17] or more subjective assessments.[14,18,19]

A recent systematic review has established the breadth of tools available to assess overall trial risk (and to use this assessment to define the monitoring strategy) in the setup stage,[20] but so far there has been no such exercise for methods to assess ongoing site-level risk once a trial has started. We conducted a scoping review[21] to identify and characterise available methods.

Our aims were 1) for trialists, to establish if any published methods were supported by adequate evidence to support implementation in routine practice and 2) for researchers in this area, to consolidate the existing evidence and point towards future developments in this growing field.

Methods

We conducted a scoping review to identify methods for using centrally held clinical trial data to assess site-level risk of deviations from Good Clinical Practice (GCP) or the trial protocol, or research misconduct, and thereby to target sites for further monitoring activity. We chose scoping review methodology as we anticipated finding a variety of methods and study types, and we wanted to characterise the extent, range and nature of research activity.[22] There is no published protocol for this scoping review.

Eligibility criteria

We defined our eligibility criteria before beginning any searches, with minor refinements (mainly to the exclusion criteria) after having piloted the search strategy.

We included reports:

- Describing methods for using centrally held data (i.e. at the clinical trials unit or other trial coordinating centre) to assess, in ongoing trials, site-level risk of protocol or GCP deviation, risk of data fabrication or research misconduct, or to target sites in some other way for corrective action based on assessed risk (regardless of whether the corrective action involved an on-site monitoring visit or not);
- With methods described in enough detail that we considered them – subjectively – reproducible;
- Either published in peer-reviewed journals or available as grey literature;
- About clinical trials, not limited to trials of Investigational Medicinal Products;
- In English.

We excluded reports:

- Published before 1996 (the year of the first version of ICH GCP Guidance, E6[R1][23]);
- About quality assurance only in the context of intervention fidelity[24] or 'rater differences'[25] for subjective trial outcome measures;
- About 'data monitoring' in general, for example data monitoring committees, unless including methods for assessing site conduct;
- About 'monitoring' in any sense other than the GCP sense, e.g. clinical monitoring;
- Focusing only on trial recruitment;
- About more efficient alternatives to standard on-site activity, for example remote source data verification;
- About site selection during trial setup;
- Featuring only opinions, or lacking enough detail allowing theoretical reproduction of methods;
- Reviewing methods presented elsewhere without any new, original methods.

Information sources and search strategies

1. Database searches

We designed search strategies for the following databases:

- PubMed

- Embase (Ovid)
- Medline (Ovid)
- Web of Science (Clarivate Analytics)
- Cinahl
- Cochrane Central
- Scopus

Full database searches took place on 23 October 2017 (run and extracted by WC). The search strategy for Medline is given in Supplementary Information. We developed our search strategy following review of systematic reviews in this area[1,26] to identify relevant search terms. We developed the final list through an iterative process in September and October 2017. The final search term combined searches around two concepts: clinical trials (using terms based on those used in a previous systematic review of monitoring methods[26]) and targeted or risk-based clinical trial monitoring. Search terms were modified as required to suit the database being used. No database filters were applied.

Both reviewers (WC, CH) imported results from the seven databases into reference management software and used in-built tools to remove duplicate entries. Both reviewers carried out initial title and abstract screening for relevant reports, producing an initial shortlist of potential papers. These were reduced through review and discussion to a final list of relevant reports for full-text review. We reviewed and discussed these, using full-text reports where possible, to agree a final list of relevant reports. Throughout the process, SS acted as third reviewer where required.

In order to ensure our results were current, this element of the search strategy was repeated on 28 August 2018. WC ran the searches and conducted the title and abstract screening. A shortlist of potentially relevant reports was shared with SS and CH; SS and WC agreed a final list of additional relevant reports from this repeated search.

2. Conference abstracts

We searched for relevant conference abstracts from the first four International Clinical Trials Methodology Conferences (occurring between 2011 and 2017) and all annual meetings of the Society for Clinical Trials since 1996 (initial searches completed on 8 December 2017).

Our methods to search conference abstracts varied depending on the year and the conference: we used website search functions where available and manual PDF searches otherwise. Keywords used for the conference abstract search were based on the key database search strategy terms. These were: 'monitor', 'supervision', 'oversight', 'risk', 'performance', 'metric', 'quality', 'fraud', 'fabrication', and 'error'.

Both WC and CH performed the abstract searches. This produced an initial shortlist of potentially relevant abstracts. A final list was agreed through discussion, with SS acting as third reviewer where required.

3. Internet searches

We conducted structured searches through Google internet search engine (searches carried out 15-19 December 2017).

Google searches were performed without limitations or use of quotes. Search terms were based on the main database search: 'Risk based monitoring', 'Risk adapted monitoring', 'Central monitoring', 'Central statistical monitoring', 'Triggered monitoring', 'Targeted monitoring', 'Performance metric', 'Site metric', 'Key risk indicator', 'Site performance', 'Centre performance', 'Detect fraud' and 'Detect fabrication'. We reviewed results on the first 20 pages; if there were no relevant results on any three consecutive pages before this, we stopped reviewing.

WC and CH conducted the searches. Any potential additions to the included list of reports were discussed and agreed, with SS acting as third reviewer where required.

4. References, citations and author contact

To identify other relevant reports, we reviewed references (manually) and citations (using Web of Science) of all papers included or considered for inclusion in the final results, and of review articles relevant to the topic. Whenever required, we contacted report authors to help ascertain if given reports should be included, and to ask about the availability of full-text articles.

Data collection

We extracted data from full journal articles, where available. We recorded data into an Excel-based tool. WC carried out the final data collection used for this report, with SS double-checking all data for inclusion; consensus was reached on any areas of

disagreement. Article authors were contacted (two attempts maximum) for missing descriptive data and further clarifications.

Our data collection template was designed and agreed prior to any data collection, with minor refinement after a first review of all relevant papers (a list of data collection variables is available as Supplementary Information). We collected descriptive data about each of the included reports, including year of publication, type of report and details of the trial(s) it was embedded in. We also looked for any information on cost implications of the proposed methods.

When designing this study, although we predicted we would find a range of methods, we agreed that most of them would in essence address a classification problem, i.e. methods to assign sites a status as 'concerning' or 'not-concerning', with a 'true' deviation status – i.e. confirmed existence of meaningful problems – that could be uncovered by further review. The 'gold standard' reference test required to assess true status might be study-specific, but could be on-site monitoring or, if the true status was created through simulation, prior knowledge.

We considered a key measure of the reported methods' effectiveness to be a demonstrated ability, ideally in a real-life setting, not only to detect 'true' sites of concern, but also to show with confidence that sites apparently not of concern are performing well. We therefore aimed to summarise the available information on classification, i.e. any or all of specificity, sensitivity, positive and negative predictive value. We gathered the best reported classification statistics for each method, or, if this was not reported, used available statistics, e.g. number of true and false positives, to calculate these. These calculations were verified

by an independent statistician at the Medical Research Council Clinical Trials Unit at University College London.

We did not formally assess the quality of the studies. However, review of the QUADAS-2 tool for quality assessment in diagnostic accuracy studies[27] informed development of our data collection template.

Synthesis of results

We did not combine results for individual studies, as it was clear through preliminary review of relevant papers that we would have a variety of study types. Instead, the evidence is summarised descriptively.

Results

Figure 1 gives a PRISMA flow diagram[28] showing the different stages of the review. From the various data sources, we ultimately included 30 reports in our final dataset. 21 of these are peer-reviewed publications. The results are characterised in **Table 1** and listed in full in **Table 2. Figure 2** shows reports by year of publication.

When specific trials were mentioned, they involved various health conditions and were in various geographical settings. Information on trial intervention was not often available, but where it was, methods had most often been used in phase III trials of investigational medicinal products (IMP). The IMP risk category,[29] when known, was either 'licensed and used within its licensed indication', or 'licensed and used outside its licensed indication' (i.e. we found no reports involving trials of unlicensed IMPs).

We classified 20/30 of our results as central statistical monitoring methods, of which 7 focussed on detection of investigator fraud or research misconduct. We classified 9, including one of the 20 that used central statistical monitoring, as 'triggered monitoring', i.e. review of each trial site against pre-set thresholds in key performance metrics, usually without any statistical testing. A final 2 we could not fit into either of these categories; these involved using measured site metrics to directly compare sites against one another.[30,31]

21/30 reports included some assessment of the effectiveness of the methods; these are summarised in **Table 3.** The most common experimental designs were to explore the methods' characteristics using real trial data with no known integrity issues (n=9), and simulating data integrity problems at sites within real trial datasets then using the method to try to identify the problem sites (n=6).

Of the 21 reports, 9 had no information about sites' 'true' status, i.e. whether the problems identified through central monitoring constitute meaningful problems (either recorded through on-site monitoring or audit activity, or known because statuses were created through simulation). One report[32] only contained case studies, i.e. partial and selective reporting. Seven[14,33–38] had partial information, e.g. some of sites' true statuses were reported, but not all. Two explored classification ability through extensive simulation,[39,40] and 2 had detailed information from a limited set of scenarios on the number of true and false positives and negatives.[19,41] The best reported or deducible classification ability for the 11 papers with at least some information on sites' 'true' status is shown in **Table 4;** in many cases this is limited by considerable amounts of missing or unclear data.

Some papers report on actual or theoretical cost savings from reduced on-site monitoring,[33,42,43] and others comment on the risk of incurring costs if their proposed central monitoring method identifies sites that do not in fact have meaningful problems (i.e. false positives).[19,37] However, no papers give detailed costings for the development, implementation and maintenance of the central monitoring methods themselves.

Discussion

We conducted a scoping review to identify published methods for assessing the risk of not taking corrective action at trial sites at a given time. Although our search looked for reports from any time after 1995, over half of our results are from after 2013, highlighting the recent growth of risk-based monitoring concepts. Host trials for these methods were most often phase III trials of licensed investigational medicinal products, i.e. somewhat lower risk compared to earlier phase trials of newer treatments.[29] Around a third of our results were not full, peer-reviewed reports, reflecting a wider problem with evidence supporting trial conduct methods, i.e. that researchers can feasibly produce posters and abstracts for conferences, but may not have time or incentive to produce comprehensive reports.[44]

Identified methods were mainly in two broad categories (with some overlap). Most were about central statistical monitoring, which uses statistical testing of all or a subset of trial data items to compare sites and identify atypical trial centres. A minority described triggered monitoring techniques, whereby sites are assessed against pre-specified site metric threshold rules (usually binary), with sites meeting the greatest number of 'triggers' being considered the most concerning. Several authors note that central statistical monitoring needs sufficient overall and per-site sample sizes for adequate statistical power.[17,19,37] Triggered monitoring, however, can be used at any stage of a trial's recruitment (especially with trigger rules based on single instances of a given protocol violation, for example). We therefore suggest use of the techniques is not mutually exclusive.

Although the reports on central statistical monitoring described a range of methods, there were some commonalities. Previous papers have reviewed these methods in more detail.[17,41,45–47]

Nearly half of the central statistical monitoring reports focused on identification of data fabrication. The possibility of fraud is a serious concern to trialists and a threat to wider trust in science.[48] It was possibly an important factor in establishing 100% on-site verification of trial data as a common monitoring approach.[49,50] This may help explain the prevalence of reports about fraud detection, as some may see the priority in risk-based monitoring to be establishing its fraud detection ability compared with 100% source data verification. However, although the incidence of data fraud is difficult to quantify, cases of extensive data fabrication appear rare enough to have individual notoriety.[51] Further, methods to detect fraud are necessarily rather selective, and therefore may not alone be suitable for trialists looking to detect more common, lower-level data integrity issues.

We collected data on how the proposed methods we identified had been evaluated. A number of reports only presented proposed, untested methods, or only selected case studies to demonstrate the methods' performance. Of those that presented more detailed evaluation, a common limitation was that the 'true' status both of identified problem sites and sites apparently not of concern was often not available, or only partially available. It was therefore difficult to know if the 'concern' status of sites in central monitoring results represented meaningful problems or not. In addition, a number of studies use simulation to create 'true' sites of concern; these raise the additional question of whether these simulations reflect real-life issues, though the involvement of clinicians (i.e. those who would provide real-life trial data) in the simulation process of some reports[19,36] is reassuring.

Clearly, a balance must be struck between minimising cost and adequately mitigating key risks. Many false positives could be expensive (e.g. additional on-site monitoring visits for no gain), but many false negatives could be catastrophic (e.g. missing serious risks to trial participants' safety). Some authors do acknowledge this, and even suggest that a high false positive rate could be acceptable in a method used to broadly target additional scrutiny at particular sites.[38]

It is important to recognise the limitations of the available 'gold-standards' in the classification of sites. When methods are tested using simulated or real-but-adjusted data, it may be difficult to know how well these accurately recreate real-life situations. When central monitoring methods are tested in real, ongoing trials, on-site monitoring may be an imperfect reference test, in that it may not be able to identify all problems. By contrast, it is clear that central monitoring, with its enhanced inter- and intra-site review, can identify issues that a single team at one site for a limited time might not.[47]

It could be argued that at least some of the methods we have identified do not need extensive evaluation because they prove their own worth. For instance, they help identify outliers that in some cases are self-evidently meaningful problems to resolve. We acknowledge that some central monitoring activities identify 'known' problems (e.g. identifying weekend visit dates, which in most cases are unlikely to be correct) and are valuable for data cleaning purposes. However, we were specifically interested in the more nuanced use of these methods to identify sites of 'concern', at which monitoring activity may be targeted, and consequently sites 'not of concern', monitoring of which may be reduced or omitted. This element does need adequate proof before wider adoption and, if it is shown to be effective, could have significant benefits. In light of the limitations we have

described here, we do not believe any methods have yet fully demonstrated that they should be adopted more widely as a means to limit on-site monitoring of sites deemed, based on review of centrally-held data, not to be of concern.

Aside from some comments on the potential cost of investigating false positive central monitoring results,[19,52] the reports we identified contained limited information on the cost of developing and implementing their methods. As well as uncertainty about how to develop relevant methods, uncertainty or concern about costs involved is a substantial barrier to adoption of risk-based monitoring.[53]

Through our various searches, we identified 24 commercial companies which, we were at least reasonably certain, had a method relevant to our search (data not shown). Only a small proportion of these had published details of their methods in peer-reviewed journals, although just over half had some detail in grey literature sources. This highlights another difficulty in disseminating new trial conduct methods, i.e. that they may be commercially sensitive and, unlike evidence about treatments being tested in trials, there is no compulsion to publish new evidence. It is somewhat contradictory that while risk-based monitoring has come about partly to reduce costs of trials, some risk-based monitoring methods are available only for a fee.

Further work is needed to fully demonstrate the effectiveness of these dynamic site risk assessment methods which, alongside pre-trial risk assessments, form the core of risk-based monitoring. We therefore recommend the following:

1. **Coordinate research efforts.** From the scoping review and contact with report authors, it was clear that various small research projects relevant to this topic

were ongoing, but mostly in isolation. Researchers in this area should take stock of existing research, and set clear priorities to ensure research time is well-spent.

2. **Standardise monitoring studies.** Core outcome sets[54] or other mechanisms to standardise studies about monitoring would improve study quality and may facilitate cross-study evidence synthesis.

3. **Share evidence.** Time, commercial sensitivity and perceived reputational risk could all be barriers to publishing evidence about monitoring practices. However, additional, publicly-available evidence to support best monitoring practice will allow trialists in all settings to adopt new methods with confidence.

4. **Publish full papers.** Conference abstracts and posters are a useful way to disseminate basic information about new ideas, but rarely have enough detail to allow replication or robustly demonstrate effectiveness. As this emerging field cannot be built on abstracts alone, we encourage researchers to publish full, peer-reviewed papers about their monitoring methods.

5. **Look to combine complementary methods.** Although work has been done on a number of distinct risk-based monitoring methods, an optimal monitoring plan might involve a combination of these, including both central statistical monitoring and triggered monitoring. A collaborative approach to combining existing methods could help develop and test such an idea.

We acknowledge several limitations. Our database searches identified relevant material from disparate locations, including abstracts in conferences in unrelated research fields. It is possible that other abstract collections include relevant material, but it was not feasible to

find and hand-search all of these. Although the internet searches made little contribution to the final list of included reports, they may have been limited by known reproducibility problems.[55]

Scoping review methodology advises that relevant experts in a field are surveyed to help identify other relevant work.[56] We have not formally done this. We have, however, contacted most authors of included reports for clarifications, and this has not highlighted any additional relevant reports.

Although we pre-specified our aims and our eligibility criteria, we added some detail to our exclusion criteria during the course of our review to help us decide about certain reports. It is possible that other researchers repeating the same review might result in a slightly different list, but we believe this might only affect the 'method-only' papers, which do not form the key part of our conclusions. The comprehensive nature of our search strategy, including review of reference and citation lists, gives us confidence that our report is a sound overview of the state of the evidence in this research area.

We have not performed a formal quality assessment of reports we found, however, this is considered by some to be unnecessary in scoping review methodology.[22] There is also no validated way to review the quality of risk-based monitoring studies, although we used the QUADAS-2 tool, designed to assess the quality of diagnostic studies, to inform our data collection template.

Finally, we acknowledge that some time has now passed since we first conducted our search for relevant evidence. Conscious of this, we repeated the main database search in 2018 (albeit with only one author conducting title and abstract screening) and added three relevant reports. We are not aware of any research published since then that might change

our overall conclusions. If evidence is now available that addresses the limitations we have highlighted in the existing literature, we would certainly consider this a positive development.

Our scoping review highlighted some promising evidence for risk-based monitoring in ongoing trials. However, currently published methods may not yet have demonstrated their efficacy or cost-effectiveness well enough for trialists to implement them with confidence as a means to target or omit on-site visits. A more coordinated, collaborative and transparent approach to developing and sharing evidence in this field, including industry and academic partners, could help it grow beyond its current nascent state, and could contribute to risk-based monitoring more quickly entering routine practice.

Acknowledgements

We would like to thank the authors of reports – some referenced in this manuscript and some ultimately not – for their time in answering our questions and discussing their work with us. We would also like to thank: systematic reviewers at Medical Research Council Clinical Trials Unit at University College London (MRC CTU at UCL) for their advice; Meredith Martyn for verifying the classification statistics, Sharon B Love for project support and Matthew R Sydes for advice, project support and review of this manuscript.

Declaration of Conflicting Interests

SS and WC are part of the TEMPER study team, one of the studies included in the final results of this paper. CH and VYE declare no conflict of interest.

Funding

This work was supported by the MRC London Hub for Trial Methodology Research (MC_UU_12023/24). The idea for this work arose from the TEMPER study, which was funded by a grant from Cancer Research UK (C1495/A13305).

Supplementary information

1. Completed PRISMA-ScR Checklist
2. Search strategy from Medline
3. Annotated list of data collection variables

References

1 Macefield RC, Beswick AD, Blazeby JM, *et al.* A systematic review of on-site monitoring methods for health-care randomised controlled trials. *Clin Trials* 2013;**10**:104–24.http://ctj.sagepub.com/content/10/1/104.abstract

2 Sheetz N, Wilson B, Benedict J, *et al.* Evaluating Source Data Verification as a Quality Control Measure in Clinical Trials. *Ther Innov Regul Sci* 2014;**48**:671–80.http://dij.sagepub.com/content/48/6/671.abstract

3 Reith C, Landray M, Devereaux PJ, *et al.* Randomized Clinical Trials — Removing Unnecessary Obstacles. *N Engl J Med* 2013;**369**:1061–5. doi:10.1056/NEJMsb1300760

4 Duley L, Antman K, Arena J, *et al.* Specific barriers to the conduct of randomized trials. *Clin Trials* 2008;**5**:40–8.http://ctj.sagepub.com/content/5/1/40.abstract

5 Olsen R, Bihlet AR, Kalakou F, *et al.* The impact of clinical trial monitoring approaches on data integrity and cost—a review of current literature. *Eur J Clin Pharmacol* 2016;**72**:399–412. doi:10.1007/s00228-015-2004-y

6 Tudur Smith C, Stocken DD, Dunn J, *et al.* The value of source data verification in a cancer clinical trial. *PLoS One* 2012;**7**:12.

7 Buyse M, Squifflet P, Coart E, *et al.* The impact of data errors on the outcome of randomized clinical trials. *Clin Trials* 2017;:174077451771615. doi:10.1177/1740774517716158

8 MRC/DH/MHRA Joint Project. Risk-adapted Approaches to the Management of Clinical Trials of Investigational Medicinal Products. 2011.https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/343677/Risk-adapted_approaches_to_the_management_of_clinical_trials_of_investigational_medicinal_products.pdf (accessed 8 Sep 2017).

9 Food and Drug Administration. Guidance for Industry: Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring. 2013.http://www.fda.gov/downloads/Drugs/.../Guidances/UCM269919.pdf (accessed 8 Sep 2017).

10 European Medicines Agency. Reflection paper on risk based quality management in clinical trials. 2013.http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/11/WC500155491.pdf (accessed 16 Jan 2017).

11 Brosteanu O, Houben P, Ihrig K, *et al.* Risk analysis and risk adapted on-site monitoring in noncommercial clinical trials. *Clin Trials* 2009;**6**:585–96.http://ctj.sagepub.com/content/6/6/585.abstract

12 Brosteanu O, Schwarz G, Houben P, *et al.* Risk-adapted monitoring is not inferior to extensive on-site monitoring: Results of the ADAMON cluster-randomised study. *Clin Trials* 2017;**14**. doi:10.1177/1740774517724165

13 Journot V, Pignon JP, Gaultier C, *et al.* Validation of a risk-assessment scale and a risk-

adapted monitoring plan for academic clinical research studies--the Pre-Optimon study. *Contemp Clin Trials* 2011;**32**:16–24.

14 Stenning SP, Cragg WJ, Joffe N, *et al.* Triggered or routine site monitoring visits for randomised controlled trials: results of TEMPER, a prospective, matched-pair study. *Clin Trials* 2018;:174077451879337. doi:10.1177/1740774518793379

15 Sullivan LB. The Current Status of Risk-Based Monitoring. http://www.appliedclinicaltrialsonline.com/current-status-risk-based-monitoring (accessed 6 Jun 2018).

16 International Conference on Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Integrated Addendum to ICH E6(R1): guideline for good clinical practice E6(R2). 2016.https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R2__Step_4_2016_1109.pdf (accessed 25 Feb 2017).

17 Kirkwood AA, Cox T, Hackshaw A. Application of methods for central statistical monitoring in clinical trials. *Clin Trials* 2013;**10**:783–806.http://ctj.sagepub.com/content/10/5/783.abstract

18 Tudur Smith C, Williamson P, Jones A, *et al.* Risk-proportionate clinical trial monitoring: an example approach from a non-commercial trials unit. *Trials* 2014;**15**:1–10. doi:10.1186/1745-6215-15-127

19 Knepper D, Lindblad AS, Sharma G, *et al.* Statistical Monitoring in Clinical Trials: Best Practices for Detecting Data Anomalies Suggestive of Fabrication or Misconduct. *Ther Innov Regul Sci* 2016;**50**:144–54. doi:10.1177/2168479016630576

20 Hurley C, Shiely F, Power J, *et al.* Risk based monitoring (RBM) tools for clinical trials: A systematic review. *Contemp Clin Trials* 2016;**51**:15–27. doi:10.1016/j.cct.2016.09.003

21 Daudt HM, van Mossel C, Scott SJ. Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. *BMC Med Res Methodol* 2013;**13**:48. doi:10.1186/1471-2288-13-48

22 Pham MT, Rajić A, Greig JD, *et al.* A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Res Synth Methods* 2014;**5**:371–85. doi:10.1002/jrsm.1123

23 International Conference on Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Guideline for good clinical practice E6(R1). 1996.http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R1_Guideline.pdf (accessed 8 Sep 2017).

24 Devito Dabbs A, Song M-K, Hawkins R, *et al.* An intervention fidelity framework for technology-based behavioral interventions. *Nurs Res* 2011;**60**:340–7. doi:10.1097/NNR.0b013e31822cc87d

25 Opler MGA, Yavorsky C, Daniel DG. Positive and Negative Syndrome Scale (PANSS) Training: Challenges, Solutions, and Future Directions. *Innov Clin Neurosci* 2017;**14**:77–81.http://www.ncbi.nlm.nih.gov/pubmed/29410941 (accessed 4 Jun 2018).

26 Bakobaki J, Joffe N, Burdett S, *et al.* A systematic search for reports of site monitoring technique comparisons in clinical trials. *Clin Trials* 2012;**9**:777–80.http://ctj.sagepub.com/content/9/6/777.abstract

27 Whiting PF, Rutjes AWS, Westwood ME, *et al.* QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med* 2011;**155**:529. doi:10.7326/0003-4819-155-8-201110180-00009

28 Moher D, Liberati A, Tetzlaff J, *et al.* Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 2009;**6**:e1000097. doi:10.1371/journal.pmed.1000097

29 Organisation for Economic Co-operation and Development. OECD Recommendation on the Governance of Clinical Trials. 2013. https://www.oecd.org/sti/sci-tech/oecd-recommendation-governance-of-clinical-trials.pdf

30 Smith A, Seltzer J, A. S. An In-Process Scaling Model: A Potential Framework for Data Monitoring Committees and Clinical Trial Quality Improvement. *Drug Inf J* 2012;**46**:8–12. doi:10.1177/0092861511427864

31 Djali S, Janssens S, Van Yper S, *et al.* How a Data-Driven Quality Management System Can Manage Compliance Risk in Clinical Trials. *Drug Inf J* 2010;**44**:359–73.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed12&NEWS=N&AN=359239259

32 Edwards P, Shakur H, Barnetson L, *et al.* Central and statistical data monitoring in the Clinical Randomisation of an Antifibrinolytic in Significant Haemorrhage (CRASH-2) trial. *Clin Trials* 2013;**11**:336–43. doi:10.1177/1740774513514145

33 Biglan KM, Brocht A, Raca P, *et al.* Implementing Risk-Based Monitoring (RBM) in STEADY-PD III, A Phase III Multi-site Clinical Drug Trial for Parkinson Disease. *Mov Disord* 2016;**31**:E10–E10.http://onlinelibrary.wiley.com/doi/10.1002/mds.26749/full

34 Knott C, Valdes-Marquez E, Landray M, *et al.* Improving efficiency of on-site monitoring in multicentre clinical trials by targeting visits. *Trials* 2015;**16**:1. doi:10.1186/1745-6215-16-s2-o49

35 Lindblad AS, Manukyan Z, Purohit-Sheth T, *et al.* Central site monitoring: Results from a test of accuracy in identifying trials and sites failing Food and Drug Administration inspection. *Clin Trials* 2014;**11**:205–17.http://ctj.sagepub.com/content/11/2/205.abstract

36 O'Kelly M. Using statistical techniques to detect fraud: a test case. *Pharm Stat* 2004;**3**:237–46. doi:10.1002/pst.137

37 van den Bor RM, Vaessen PW, Oosterman BJ, *et al.* A Computationally Simply Central Monitoring Procedure was Proposed and Effectively Applied to Empirical Trial data with Known Fraud. *J Clin Epidemiol* 2017;**0**. doi:10.1016/j.jclinepi.2017.03.018

38 Wu X, Carlsson M. Detecting data fabrication in clinical trials from cluster analysis perspective. *Pharm Stat* 2011;**10**:257–64. doi:10.1002/pst.462

39 Desmet L, Venet D, Doffagne E, *et al.* Linear mixed-effects models for central statistical monitoring of multicenter clinical trials. *Stat Med* 2014;**33**:5265–79. doi:10.1002/sim.6294

40 Desmet L, Venet D, Doffagne E, *et al.* Use of the Beta-Binomial Model for Central Statistical Monitoring of Multicenter Clinical Trials. *Stat Biopharm Res* 2017;**9**:1–11. doi:10.1080/19466315.2016.1164751

41 Pogue JM, Devereaux P, Thorlund K, *et al.* Central statistical monitoring: Detecting fraud in clinical trials. *Clin Trials* 2013;**10**:225–35. doi:10.1177/1740774512469312

42 Diani CA, Rock A, Moll P. An evaluation of the effectiveness of a risk-based monitoring approach implemented with clinical trials involving implantable cardiac medical devices. *Clin Trials* 2017;:1740774517723589. doi:10.1177/1740774517723589

43 Agrafiotis DK, Lobanov VS, Farnum MA, *et al.* Risk-based Monitoring of Clinical Trials: An Integrative Approach. *Clin Ther* 2018;**40**:1204–12. doi:10.1016/j.clinthera.2018.04.020

44 Brueton V, Tierney JF, Stenning S, *et al.* Identifying additional studies for a systematic review of retention strategies in randomised controlled trials: making contact with trials units and trial methodologists. *Syst Rev* 2017;**6**:167. doi:10.1186/s13643-017-0549-9

45 Oba K. Statistical challenges for central monitoring in clinical trials: a review. *Int J Clin Oncol* 2016;**21**:28–37.

46 Timmermans C, Venet D, Burzykowski T. Data-driven risk identification in phase III clinical trials using central statistical monitoring. *Int J Clin Oncol* 2016;**21**:38–45. doi:10.1007/s10147-015-0877-5

47 Venet D, Doffagne E, Burzykowski T, *et al.* A statistical approach to central monitoring of data quality in clinical trials. *Clin Trials* 2012;**9**:705–13. doi:10.1177/1740774512447898

48 Buyse M, Evans SJW, Buyse M, *et al.* Fraud in Clinical Trials. In: *Encyclopedia of Biostatistics*. Chichester, UK: : John Wiley & Sons, Ltd 2005. doi:10.1002/0470011815.b2a01026

49 Peto R, Collins R, Sackett D, *et al.* The trials of Dr. Bernard Fisher: A European perspective on an American episode. *Control Clin Trials* 1997;**18**:1–13. doi:10.1016/S0197-2456(96)00225-5

50 Seachrist L. Scientific misconduct. NIH tightens clinical trials monitoring. *Science (80-)* 1994;**264**:499–499. doi:10.1126/science.8160006

51 George SL, Buyse M. Data fraud in clinical trials. *Clin Invest (Lond)* 2015;**5**:161–73. doi:10.4155/CLI.14.116

52 van den Bor RM, Vaessen PWJ, Oosterman BJ, *et al.* A computationally simple central monitoring procedure, effectively applied to empirical trial data with known fraud. *J Clin Epidemiol* 2017;**87**:59–69. doi:10.1016/j.jclinepi.2017.03.018

53 Hurley C, Sinnott C, Clarke M, *et al.* Perceived barriers and facilitators to Risk Based Monitoring in academic-led clinical trials: a mixed methods study. *Trials* 2017;**18**:423. doi:10.1186/s13063-017-2148-4

54 Williamson PR, Altman DG, Blazeby JM, *et al.* Developing core outcome sets for clinical trials: Issues to consider. *Trials* 2012;**13**:132. doi:10.1186/1745-6215-13-132

55 Ćurković M, Košec A. Bubble effect: including internet search engines in systematic reviews introduces selection bias and impedes scientific reproducibility. *BMC Med Res Methodol* 2018;**18**:130. doi:10.1186/s12874-018-0599-2

56 Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005;**8**:19–32. doi:10.1080/1364557032000119616

57 Almukhtar T, Glassman A. Monitoring of adverse events reporting in multicenter clinical trials using a mixed effect regression model. In: *36th annual meeting of the Society for Clinical Trials*. 2015. P45.

58 Atanu B, Savanur S, Suman K. Risk based monitoring in clinical trial: an application with neural networking. *Biostat Biometrics* 2017;**3**:555624.

59 Bailey L, Brudenell Straw FK, George SE. Implementing a risk based monitoring approach in the early phase myeloma portfolio at Leeds CTRU. *Trials* 2017;**18(Suppl 1**:220.

60 Bengtsson SKS. *Risk based monitoring in clinical studies - improving data quality*. 2017.https://lup.lub.lu.se/student-papers/search/publication/8928535

61 Dress J, U. H, C. W, *et al.* High quality risk management for clinical trials: Use the data at your hands to manage risk in your clinical trials. *Clin Trials* 2011;**8**:469. doi:http://dx.doi.org/10.1177/1740774511413037

62 Kodama A, Cavalcanti A, Buehler A, *et al.* Application of simple statistical methods to evaluated possible data fabrication: An example using the act trial. *Clin Trials* 2010;**7**:445. doi:10.1177/1740774510374795

63 Taylor RN, McEntegart DJ, Stillman EC, *et al.* Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Inf J* 2002;**36**:115–25.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed8&NEWS=N&AN=34208207

64 Timmermans C, Doffagne E, Venet D, *et al.* Statistical monitoring of data quality and consistency in the Stomach Cancer Adjuvant Multi-institutional Trial Group Trial. *Gastric Cancer* 2016;**19**:24–30. doi:10.1007/s10120-015-0533-9

65 Valdez-Marquez E, H.C. J, L. M. A key risk indicator approach to central statistical monitoring in multicentre clinical trials: Method development in the context of an ongoing large-scale randomized trial. *Trials* 2011;**12**. doi:http://dx.doi.org/10.1186/1745-6215-12-S1-A135

66 Valdes-Marquez E, J.C. H, J. A. Central statistical monitoring in multicentre clinical trials: Developing statistical approaches for analysing key risk indicators. *Trials* 2013;**14**:267.

67 Whitham D, Turzanski J, Bradshaw L, *et al.* Development of a standardised set of metrics for monitoring site performance in multicentre randomised trials: a Delphi study. *Trials* 2018;**19**:557. doi:10.1186/s13063-018-2940-9

68 Zink RC, Dmitrienko A, Dmitrienko A. Rethinking the Clinically Based Thresholds of TransCelerate BioPharma for Risk-Based Monitoring. *Ther Innov Regul Sci* 2018;**52**:560–71. doi:10.1177/2168479017738981

1

Tables

Table 1: general characteristics of included studies

Characteristic	Number (total=30)	Percentage (%)
Publication year		
1996-2000	0	0%
2000-2005	2	7%
2006-2010	2	7%
2010-2015	13	43%
2016-2018	13	43%
Type of source		
Peer-reviewed paper	21	70%
Conference abstract or poster	8	27%
Thesis	1	3%
Disease setting of trial involved		
Cardiovascular disease	4	13%
Emergency medicine	1	3%
Haematology	1	3%
Infectious diseases	1	3%
Mental health	3	10%
Neurology	1	3%
Oncology	3	10%
Ophthalmology	1	3%
Renal disease	1	3%
Respiratory disease	1	3%
Not known or no specific trial involved	13	43%
Geographical setting of trials involved		
Brazil	1	3%
International	7	23%
Japan	1	3%
North America	4	13%
UK	2	7%
Not known or no specific trial involved	15	50%
Use of Investigational Medicinal Product (IMP) in involved trials		
Involves IMP	14	47%
No IMP	1	3%
Not known or no specific trial involved	15	50%
Phase of trials involved		
Phase I	0	0%
Phase II	1	3%
Phase II and III	1	3%
Phase III	9	30%
Not known or no specific trial involved	19	63%
Status of investigational medicinal product used[a]		
Unlicensed	0	0%
Licensed, used outside of its licensed indication	5	17%
Licensed, used within its licensed indication	4	13%
Not known or no specific trial involved	22	73%

Focus of work[a]		
Central statistical monitoring, focus on fraud or misconduct	7	23%
Central statistical monitoring, general	13	43%
Triggered monitoring	9	30%
Other method(s) for highlighting sites at risk	2	7%
Scope of work		
Description or development of method	9	30%
Some assessment of methods' effectiveness	21	70%

1

2 [a] Categories not mutually exclusive

Table 2: full listing of all included reports

First author, publication year	Type of source	Focus of work	Scope of work
Agrafiotis, 2018[43]	Peer-reviewed paper	Triggered monitoring	Some assessment of methods' effectiveness
Almukhtar, 2015[57]	Conference abstract/poster	Central statistical monitoring, general	Description or development of method
Atanu, 2017[58]	Peer-reviewed paper	Central statistical monitoring, general	Description or development of method
Bailey, 2017[59]	Conference abstract/poster	Triggered monitoring	Description or development of method
Bengtsson, 2017[60]	Thesis	Central statistical monitoring, general	Some assessment of methods' effectiveness
Biglan, 2016[33]	Conference abstract/poster	Triggered monitoring	Some assessment of methods' effectiveness
Desmet, 2014[39]	Peer-reviewed paper	Central statistical monitoring, general	Some assessment of methods' effectiveness
Desmet, 2017[40]	Peer-reviewed paper	Central statistical monitoring, general	Some assessment of methods' effectiveness
Diani, 2017[42]	Peer-reviewed paper	Triggered monitoring	Some assessment of methods' effectiveness
Djali, 2010[31]	Peer-reviewed paper	Other method(s) for highlighting sites at risk (combines site metric scores directly to flag sites of concern)	Some assessment of methods' effectiveness
Dress, 2011[61]	Conference abstract/poster	Triggered monitoring	Description or development of method
Edwards, 2014[32]	Peer-reviewed paper	Central statistical monitoring with triggered monitoring	Some assessment of methods' effectiveness
Kirkwood, 2013[17]	Peer-reviewed paper	Central statistical monitoring, general	Some assessment of methods' effectiveness
Knepper, 2016[19]	Peer-reviewed paper	Central statistical monitoring, focus on fraud or misconduct	Some assessment of methods' effectiveness
Knott, 2015[34]	Conference abstract/poster	Central statistical monitoring, general	Some assessment of methods' effectiveness
Kodama, 2010[62]	Conference abstract/poster	Central statistical monitoring, focus on fraud or misconduct	Some assessment of methods' effectiveness

First author, publication year	Type of source	Focus of work	Scope of work
Lindblad, 2014[35]	Peer-reviewed paper	Central statistical monitoring, general	Some assessment of methods' effectiveness
O'Kelly, 2004[36]	Peer-reviewed paper	Central statistical monitoring, focus on fraud or misconduct	Some assessment of methods' effectiveness
Pogue, 2013[41]	Peer-reviewed paper	Central statistical monitoring, focus on fraud or misconduct	Some assessment of methods' effectiveness
Smith, 2012[30]	Peer-reviewed paper	Other method(s) for highlighting sites at risk (use of "statistical process control methodology" to combine per-site risk indicator scores)	Description or development of method
Stenning, 2018[14]	Peer-reviewed paper	Triggered monitoring	Some assessment of methods' effectiveness
Taylor, 2002[63]	Peer-reviewed paper	Central statistical monitoring, focus on fraud or misconduct	Some assessment of methods' effectiveness
Timmermans, 2016[64]	Peer-reviewed paper	Central statistical monitoring, general	Some assessment of methods' effectiveness
Tudur Smith, 2014[18]	Peer-reviewed paper	Triggered monitoring	Description or development of method
Valdez-Marquez, 2011[65]	Conference abstract/poster	Central statistical monitoring, general	Description or development of method
Valdez-Marquez, 2013[66]	Conference abstract/poster	Central statistical monitoring, general	Description or development of method
van den Bor, 2017[52]	Peer-reviewed paper	Central statistical monitoring, focus on fraud or misconduct	Some assessment of methods' effectiveness
Whitham, 2018[67]	Peer-reviewed paper	Triggered monitoring	Description or development of method
Wu, 2011[38]	Peer-reviewed paper	Central statistical monitoring, focus on fraud or misconduct	Some assessment of methods' effectiveness
Zink, 2018[68]	Peer-reviewed paper	Central statistical monitoring, general	Some assessment of methods' effectiveness

Table 3: Types of assessments and evidence presented by reports that included some assessments of their methods' effectiveness

First author	Case studies	Illustration of method(s) on data with no known issues	Assessment of methods' ability to identify simulated problem sites	Assessment of methods' ability to identify known problems in real trial data	Methods used in ongoing trial, results of on-site monitoring reported	Methods used in ongoing trial, effects reported on trial in general (e.g. in terms of cost or data quality)	Prospectively designed, controlled study to assess methods' ability to target on-site monitoring visits to most problematic sites
Agrafiotis[43]					X	X	
Bengtsson[60]		X					
Biglan[33]					X	X	
Desmet (2014)[39]	X	X	X				
Desmet (2017)[40]		X	X				
Diani[42]						X	
Djali[31]	X						
Edwards[32]	X						
Kirkwood[17]		X	X				
Knepper[19]			X				
Knott[34]					X		
Kodama[62]		X					
Lindblad[35]				X			
O'Kelly[36]			X				
Pogue[41]		X		X			
Stenning[14]							X
Taylor[63]		X					
Timmermans[64]		X					
van den Bor[52]				X			
Wu[38]			X	X			
Zink[68]		X					
Total	**3**	**9**	**6**	**4**	**3**	**3**	**1**

Table 4: best reported information on methods' classification ability, where available or deducible

First author	Available information on methods' classification abilities	Definition of 'positive' centres	'True' test status: real or simulated?	Test for 'true' centre status	Sensitivity[a]	Specificity[b]	Positive Predictive Value[c]	Negative Predictive Value[d]
Biglan[33]	Partial ('true' status known for only one centre; total number of centres not known)	Not clearly defined[e]	Real	On-site monitoring	Unavailable due to limited data; report states that one 'low-risk' centre was visited and considered to be misclassified (i.e. should have been 'medium risk' or 'high risk'). However, total number of sites classified and visited (overall and within each risk category) is not known.			
Desmet (2014)[39]	Explored through simulation	Presence of atypical data	Simulated	Known because simulated	Dependent on simulation scenario; no specific figure given			
	Detailed information (vital signs data used as illustrative example)	Presence of atypical data	Real	Unclear ('closer inspection')	**Reported:** 83% (10/(10+2))	**Reported:** 99% (204/(204+2))	**Calculated:** 83% (10/(10+2))	**Calculated:** 99% (204/(204+2))
Desmet (2017)[40]	Explored through simulation	Presence of atypical data	Simulated	Known because simulated	**Reported:** dependent on simulation scenario; no specific figure given	**Reported:** median specificity varied from 98%-100% depending on scenario	Not reported and not possible to calculate (results of many simulations presented)	

First author	Available information on methods' classification abilities	Definition of 'positive' centres	'True' test status: real or simulated?	Test for 'true' centre status	Sensitivity[a]	Specificity[b]	Positive Predictive Value[c]	Negative Predictive Value[d]
Knepper[19]	Detailed information	Presence of fabricated data	Simulated with physician input	Known because simulated	**Reported:** best result from 4 scenarios (study 1): 86% (6/(6+1))	**Reported:** best result from 4 scenarios (study 1a): 87% (148/(148+23)) [f]	**Reported:** best result from 4 scenarios (study 2a): 27% (3/3+8)	**Reported:** best result from 4 scenarios (study 1): 99% (132/132+1)
Knott[34]	Partial (total number of sites not reported but likely more than number whose results reported; 'true' status of any unreported centres not known)	Presence of any findings	Real	On-site monitoring	**Calculated:** 85% (11/(11+2))	**Calculated:** 88% (7/(7+1))	**Calculated:** 92% (11/(11+1)	**Calculated:** 78% (7/(7+2))
		Presence of findings 'indicative of sloppy practice' (clearer definition not reported)	Real	On-site monitoring	**Calculated:** 83% (10/((10+2))	**Calculated:** 78% (7/(7+2))	**Calculated:** 83% (10/(10+2))	**Calculated:** 78% (7/(7+2))
		Presence of serious findings (clearer definition not reported)	Real	On-site monitoring	**Calculated:** 100% (1/1+0)	**Calculated** 45% (9/(9+11))	**Calculated** 8% (1/(1+11))	**Calculated** 100% (9/(9+0))
Lindblad[35]	Partial ('true' status known only at 21/413 centres)	Presence of serious problems	Real	Regulatory inspection	**Reported:** 83% (5/((5+1))	Cannot be calculated without making assumptions about the 392/413 sites with unknown 'true' status		
		Presence of minor problems	Real	Regulatory inspection	**Reported:** 89% (8/(8+1))			
		Presence of any problems	Real	Regulatory inspection	**Reported:** 87% (13/(13+2))			

First author	Available information on methods' classification abilities	Definition of 'positive' centres	'True' test status: real or simulated?	Test for 'true' centre status	Sensitivity[a]	Specificity[b]	Positive Predictive Value[c]	Negative Predictive Value[d]
O'Kelly[36]	Detailed information, but sample of data from trial [g]	Presence of fabricated data	Simulated with physician input	Known because simulated	**Calculated:** 33% (1/(1+2))	**Calculated:** 95% (18/(18+1))	**Calculated:** 50% (1/(1+1))	**Calculated:** 90% (18/(18+2))
Pogue[41]	POISE trial data: detailed information from all sites with >= 20 randomisations	Presence of fabricated data	Real	On-site monitoring	**Reported:** different models and different thresholds give different pros and cons in terms of classification. Models 1, 3 and 5 all have at least some scenarios where both specificity and sensitivity > 80%. (Model 1 and 5, risk score >=7; Model 3, risk score >=5)			
	HOPE trial data: summary information from all sites with >= 20 randomisations	Presence of fabricated data	Real	On-site monitoring	N/a (no true positives)	**Reported:** model 1: 99% (178/(178+2))	**Calculated:** all models: 0% (no true positives, so any positives are false)	**Calculated:** all models: 100% (no true positives, so all negatives are true negatives)
Stenning[14]	Partial (only sample of negative-testing sites visited, although the study design aimed to control for this)	Presence of ≥ 1 serious (Major or Critical) finding	Real	On-site monitoring	**Calculated:** primary analysis: 52% (37/(37+34))	**Calculated:** primary analysis: 62% (8/(8+5))	**Reported:** primary analysis: 88% (37/(37+5))	**Calculated:** primary analysis: 19% (8/(8+34))
					Calculated: secondary analysis excluding re-consent findings: 59% (36/(36+25))	**Calculated:** secondary analysis excluding re-consent findings: 74% (17/(17+6))	**Reported:** secondary analysis excluding re-consent findings: 86% (36/(36+6))	**Calculated:** secondary analysis excluding re-consent findings: 40% (17/(17+25))

First author	Available information on methods' classification abilities	Definition of 'positive' centres	'True' test status: real or simulated?	Test for 'true' centre status	Sensitivity[a]	Specificity[b]	Positive Predictive Value[c]	Negative Predictive Value[d]
					Calculated: secondary analysis excluding all consent findings: 60% (29/(29+19))	**Calculated:** secondary analysis excluding all consent findings: 64% (23/(23+13))	**Reported:** secondary analysis excluding all consent findings: 69% (29/(29+13))	**Calculated:** secondary analysis excluding all consent findings: 55% (23/(23+19))
van den Bor[37]	Partial in paper, but authors confirmed that trial implemented source data verification for all sites (personal communication)	Presence of fabricated data	Real	On-site monitoring	Various situations presented, with different implications for classification ability. Median false positives below 10% for all scenarios, lower with higher m-constant; in various situations (combinations of specific m-constants with specific scenarios), the fraudulent centre is flagged >=3 times (authors' proposed threshold) 100% of the time. Some scenarios have 100% highlighting of fraudulent centre and very low false positive rate - e.g. scenario 1, m=20, scenario 2, m=2-, scenario 3, m=20 (all with false positive rate of 2%)			
Wu[38]	Partial (15/17 sites have unknown 'true' status)	Presence of fabricated data	Real	Auditing	Results presented narratively via a number of scenarios. For 'angular clustering', fourth scenario (correlation 0.7, 3 outliers) results in sensitivity, specificity, positive and negative predictive values all ≥ 98%. For 'neighborhood clustering', specificity in all scenarios is ≥ 94% and second scenario (variances 0.45, 3 outliers, cluster size 27) results in sensitivity, specificity, positive and negative predictive values all ≥ 50%.			

[a] Number of correctly flagged problem sites / (number of correctly flagged problem sites + sites incorrectly *not* flagged as concerning); thick border used to highlight results more than or equal to 90%.

[b] Number of sites correctly flagged as not concerning / (number of sites correctly flagged as not concerning + sites incorrectly flagged as concerning); thick border used to highlight results more than or equal to 90%.

[c] Number of correctly flagged problem sites / (number of correctly flagged problem sites + sites incorrectly flagged as concerning); thick border used to highlight results more than or equal to 90%.

[d] Number of sites correctly flagged as not concerning / (number of sites correctly flagged as not concerning + sites incorrectly *not* flagged as concerning); thick border used to highlight results more than or equal to 90%.

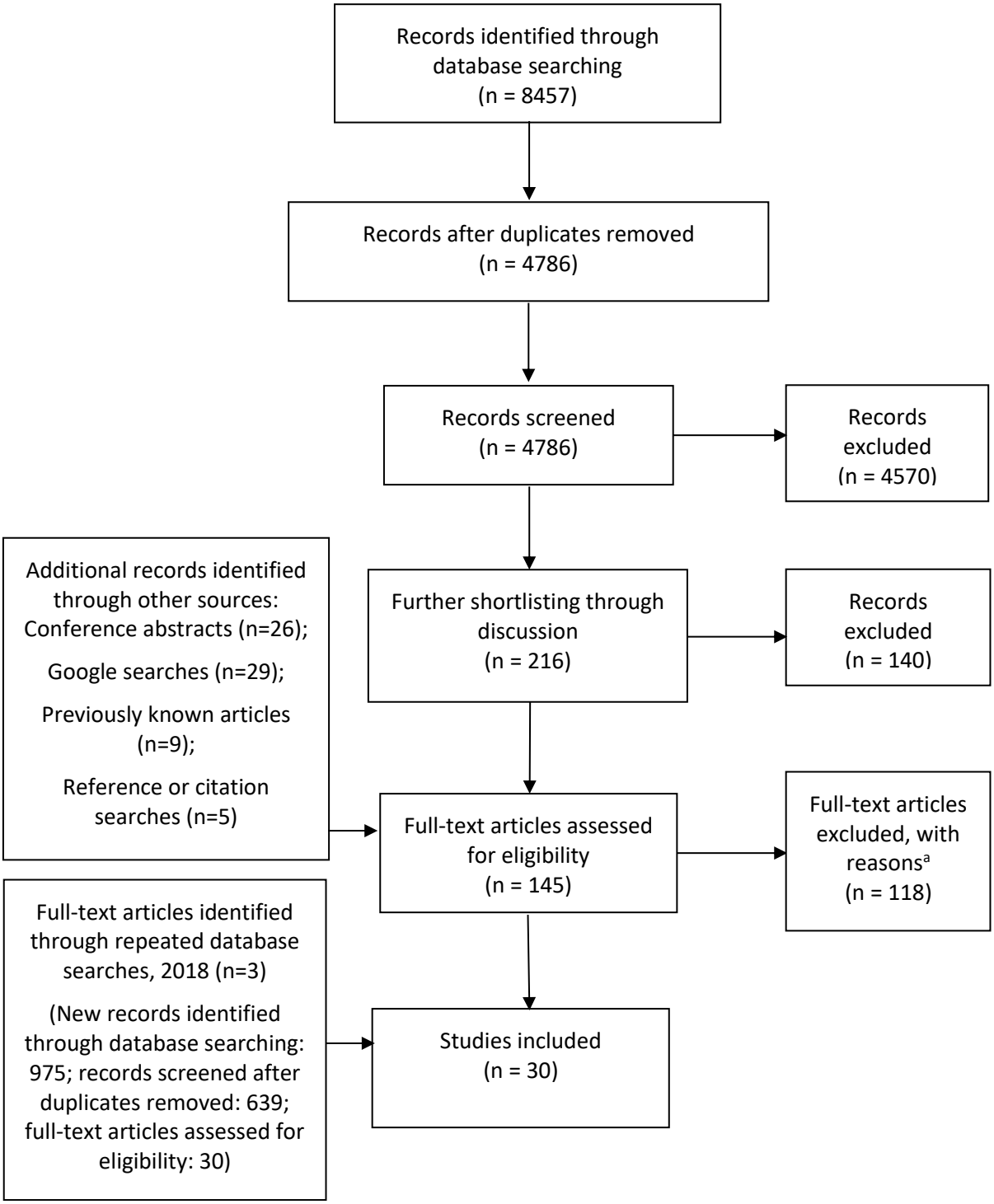
[e] One 'positive' centre is described as "reveal[ing that] RBM was not assessing risk sufficiently to drive monitoring decisions"

[f] Publication incorrectly rounds this to 86%.

[g] Approximately one third of sites included from a trial; also some uncertainty about total number of sites (sometimes reported as 21, sometimes 22; used 22 for calculations given here as this is figure in Results section)

Figures

Figure 1 – PRISMA flow diagram



[a] Reasons: no relevant methods presented (n=28); no novel methods presented (e.g. review article; n=28); method to measure variation between trial sites but no 'flagging' of sites of concern (n=25); abstract only and not enough detail to confirm relevance (n=10); duplicate or abstract where full paper also available (n=8); grey literature not considered to present reproducible methods (n=5); not about 'monitoring' according to ICH Good Clinical Practice definition (n=5); trial-level assessment only, not site-level (n=4); focus on consistency of outcome assessment only (n=4); method from observational study only, not clinical trial (n=1).

Figure 2 – Publications by year and type