Innovations in monitoring and data quality control in clinical trials

Rutger M. van den Bor

Innovations in monitoring and data quality control in clinical trials

PhD thesis. Julius center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands. Julius Clinical Research Ltd., Zeist, the Netherlands.

ISBN	978-90-393-6925-8
Author	Rutger M. van den Bor
Printed by	GVO drukkers & vormgevers B.V.

The studies presented in chapters 2, 3, 4 and 5 in this thesis were funded by Julius Clinical Research Ltd., Zeist, the Netherlands. The work leading to the study presented in chapter 6 has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no [115546], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007/2013), and EFPIA companies in kind contribution. Financial support by the Julius center for Health Sciences and Primary Care for the publication of this thesis is gratefully acknowledged.

Innovations in monitoring and data quality control in clinical trials

Innovaties in het monitoren en het bevorderen van datakwaliteit in klinische studies

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op

donderdag 11 januari 2018
des middags te 4.15 uur

door

Rutger Matthijs van den Bor

geboren op 30 augustus 1988
te Ermelo

Promotoren:	Prof. dr. C.B. Roes
	Prof. dr. D.E. Grobbee
Copromotor:	Dr. B.J. Oosterman

Manuscripts based on the studies presented in this thesis

Chapter 2 Van den Bor RM, Oosterman BJ, Oostendorp MB, Grobbee DE, Roes KCB. Efficient source data verification using statistical acceptance sampling: A simulation study. Therapeutic Innovation & Regulatory Science, 2016; 50(1):82-90.

Chapter 3 Van den Bor RM, Vaessen PWJ, Oosterman BJ, Zuithoff NPA, Grobbee DE, Roes KCB. A computationally simple central monitoring procedure, effectively applied to empirical trial data with known fraud. Journal of Clinical Epidemiology, 2017; 87:59-69.

Chapter 4 Van den Bor RM, Grobbee DE, Oosterman BJ, Vaessen PWJ, Roes KCB. Predicting enrollment performance of investigational centers in phase III multi-center clinical trials. Contemporary Clinical Trial Communications, 2017; 7:208-216.

Chapter 5 Rutger M van den Bor RM, Roes KCB, Vaessen PWJ, Oosterman BO, Grobbee DE. The impact of outcome misclassification and the value of adjudication in multi-center clinical trials. *Submitted.*

Chapter 6 Irving EA, Van den Bor RM, Welsing P, Walsh V, Alfonso-Cristancho R, Harvey C, Garman N, Grobbee DE, GetReal work package 3. Collecting and reporting safety data and monitoring trial conduct in pragmatic trials. Journal of Clinical Epidemiology, 2017 [Epub ahead of print].

Contents

Chapter 1

General introduction

Clinical trials are of key importance to understand the causal effects of medicinal treatments or therapies. Over time, however, clinical trials have become more complex and more costly [1-3]. As a consequence, it is becoming ever more necessary to critically reflect on how resources are being spent during the conduct of trials.

According to many (see e.g. [1, 4, 5, 6]), an important source of inefficiency in this respect concerns 'trial monitoring'. In the context of clinical trials, the term 'monitoring' is used for a range of activities, including e.g. the continuous assessment of the safety/benefit profile of the treatment under study by a data and safety monitoring board [7]. In discussions regarding the efficiency of clinical trials, however, it usually refers to the monitoring of investigational centers. This form of monitoring is the responsibility of one or more Clinical Research Associates (CRAs), who are appointed by the trial sponsor (typically, a pharmaceutical company) to assess whether the investigators (i.e. the treating physicians or nurses) and their staff members perform their tasks according to the trial protocol. The reason that this type of monitoring is considered inefficient is that, traditionally, CRAs frequently physically visit the centers to verify all submitted data against the source data, i.e. to perform 100% Source Data Verification (SDV). SDV, however, is very time-consuming and therefore costly [4, 8-12], while its impact on the validity of the trial results is most likely limited [13, 14].

Therefore, it is generally acknowledged, among industry parties (e.g. [5]), academic parties (e.g. [17, 18]), and regulatory agencies [1, 6], that there is a need to re-evaluate standard monitoring practices in clinical trials. These considerations are reflected in the 2015 addendum to the guidelines for Good Clinical Practice as well, in which it is stated that *"the sponsor should develop a systematic, prioritized, risk-based approach to monitoring clinical trials"* and that *"the flexibility in the extent and nature of monitoring described in this section is intended to permit varied approaches that improve the effectiveness and efficiency of monitoring"* [15, p. 39].

Objectives

As a result of the discussion, many alternative, supposedly more efficient, monitoring methodologies have been proposed (see e.g. [16] for an overview). However, a detailed

assessment of their impact on the trial results and costs is often lacking. The general aim of this thesis is to evaluate the impact of such alternative procedures, to reflect on their application, and to further their development.

Outline of this thesis

Chapter 2 focuses on one possible alternative to 100% SDV, namely the utilization of statistical sampling methodology. Although the Guidelines for Good Clinical Practice explicitly state that *"statistically controlled sampling may be an acceptable method for selecting the data to be verified"* [15, p. 39], there is limited guidance on how to actually implement statistical sampling methodology in this context, and what the impact of doing so will be. In this chapter, we propose a flexible algorithm that can be used (as such or in adapted form) for this purpose and assess its impact as compared to 100% SDV by means of a simulation.

In Chapter 3, we focus on the topic of Central Statistical Monitoring (CSM). With trial data becoming more and more available digitally and in real-time, a natural question is whether we can use these data to monitor investigational centers centrally. Doing so allows for center-by-center comparisons, which may be highly effective for the detection of deviating data patterns, e.g. those that result from fraud or data fabrication [1]. In this chapter, we propose a procedure intended to identify fraudulent investigators at an early stage, using accumulating, centrally available, subject data. The procedure is illustrated on empirical trial data with known fraud.

In Chapter 4, we investigate whether there is a relation between various characteristics of the center (e.g. the type of center, the experience of the investigator, etc.) and the performance of a center in terms of enrollment of trial subjects. Issues related to recruitment of patients in the trial are common. On the operational level, such issues may be mitigated by careful site selection and by allocating monitoring or training resources proportionally to the anticipated risk of poor enrollment. However, how to determine which centers will and which will not meet their recruitment targets a priori remains unclear. We therefore try to determine if we

can identify factors associated with center-level recruitment performance, using empirical trial data.

In Chapter 5, we focus on a monitoring procedure known as ‘adjudication’. Adjudication refers to the independent expert review of diagnoses of important clinical events, in order to detect and correct potential cases of misdiagnosis. Although common, the results of two meta-analyses (comparing pre- and post-adjudication data) suggest that the impact of adjudication on treatment effect estimates may not justify the substantial effort that is required [19, 20]. Through a series of illustrative simulations, we show the impact of random (‘non-differential’) misdiagnosis of trial endpoints on the outcomes of a trial.

Chapter 6 outlines considerations for implementing a risk-based approach to monitoring in the specific context of pragmatic trials.

Chapter 7 provides a general discussion on the topic of clinical trial monitoring, in which we (1) describe in more detail the arguments and findings that have been published on the topic, (2) critically reflect on proposed alternatives and recent developments, and (3) indicate what we believe should be directions for future research.

References

1. FDA. Guidance for industry: oversight of clinical investigations - A risk-based approach to monitoring. 2013. URL: https://www.fda.gov/RegulatoryInformation/Guidances/ucm122046.htm.
2. Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? Nature Reviews Drug Discovery 2017, 16:381-382. DOI: 10.1038/nrd.2017.70.
3. Rosenblatt M. The Changing Face of Clinical Trials - The Large Pharmaceutical Company Perspective. New England Journal of Medicine 2017; 376:52-60. DOI: 10.1056/NEJMra1510069.
4. Institute of Medicine. Assuring data quality and validity in clinical trials for regulatory decision making - Workshop report. Washington, DC: The National Academies Press. 1999. DOI: 10.17226/9623.
5. TransCelerate Biopharma Inc. Position Paper: Risk-Based Monitoring Methodology. 2013. URL: http://www.transceleratebiopharmainc.com/assets/rbm-assets/.
6. EMA. Reflection paper on risk based quality management in clinical trials. 2013. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/11/WC500155491.pdf.
7. FDA. Guidance for clinical trial sponsors – Establishment and operation of clinical trial data monitoring committees. 2006. URL: https://www.fda.gov/RegulatoryInformation/Guidances/ucm122046.htm
8. Eisenstein EL, Lemons PW II, Tardiff BE, Schulman KA, King Jolly M, Califf RM. Reducing the costs of phase III cardiovascular clinical trials. American Heart Journal 2005;149(3):482-488.
9. Califf RM. Clinical trials bureaucracy: unintended consequences of well-intentioned policy. Clinical Trials 2006; 3:496-502. DOI: 10.1177/1740774506073173.
10. Duley L, Antman K, Arena J, Avezum A, Blumenthal M, Bosch J, Chrolavicius S, Li T, Ounpuu S, Perez AC, Sleight P, Svard R, Temple R, Tsouderous Y, Yunis C, Yusuf S. Specific barriers to the conduct of randomized trials. Clinical Trials 2008; 5:40-48. DOI: 10.1177/1740774507087704.

11. Funning S, Grahnén A, Eriksson K, Kettis-Linblad Å. Quality assurance within the scope of Good Clinical Practice (GCP) - What is the cost of GCP-related activities? A survey within the Swedish Association of the Pharmaceutical Industry (LIF)'s members. Quality Assurance Journal 2009; 12:3-7. DOI: 10.1002/qaj.433.
12. Buyse M. Centralized statistical monitoring as a way to improve the quality of clinical data. Applied Clinical Trials 2014. Mar. URL: http://www.appliedclinicaltrialsonline.com/centralized-statistical-monitoring-way-improve-quality-clinical-data.
13. Sheetz N, Wilson B, Benedict J, Huffman E, Lawton A, Travers M, Nadolny P, Young S, Given K, Florin L. Evaluating source data verification as a quality control measure in clinical trials. Therapeutic Innovation and Regulatory Science 2014; 48(6):671-680. DOI: 10.1177/2168479014554400.
14. Olsen R, Bihlet AR, Kalakou F, Andersen JR. The impact of clinical trial monitoring approaches on data integrity and cost - a review of current literature. European Journal of Clinical Pharmacology 2016; 72:399-412. DOI: 10.1007/s00228-015-2004-y.
15. ICH. Guideline for good clinical practice E6(R2), step 5. 2016. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002874.pdf.
16. Hurley C, Shiely F, Power J, Clarke M, Eustace JA, Flanagan E, Kearney PM. Risk based monitoring (RBM) tools for clinical trials: A systematic review. Contemporary Clinical Trials 2016; 51:15-27. DOI: 10.1016/j.cct.2016.09.003.
17. Brosteanu O, Schwarz G, Houben P, Paulus U, Strenge-Hesse A, Zettelmeyer U, Schneider A, Hasenclever D. Risk-adapted monitoring is not inferior to extensive on-site monitoring: Results of the ADAMON cluster-randomised study. Clinical Trials 2017; Epub ahead of print. DOI: 10.1177/1740774517724165.
18. Journot V, Pignon JP, Gaultier C, Daurat V, Bouxin-Métro A, Giraudeau B, Preux PM, Tréluyer JM, Chevret S, Plättner V, Thalamas C, Clisant S, Ravaud P, Chêne G; Optimon Collaborative Group. Validation of a risk-assessment scale and a risk-adapted monitoring plan for academic clinical research studies - the Pre-Optimon study. Contemporary Clinical Trials 2011; 32(1):16-24. DOI: 10.1016/j.cct.2010.10.001.

19. Pogue J, Walter SD, Yusuf S. Evaluating the benefit of event adjudication of cardiovascular outcomes in large simple RCTs. Clinical Trials 2009; 6:239-251. DOI: 10.1177/1740774509105223.
20. Ndounga Diakou LA, Trinquart L, Hróbjartsson A, et al. Comparison of central adjudication of outcomes and onsite outcome assessment on treatment effect estimates. Cochrane Database of Systematic Reviews 2016; 3:MR000043. DOI: 10.1002/14651858.MR000043.pub2.

Chapter 2

Efficient source data verification using statistical acceptance sampling: A simulation study

Rutger M van den Bor
Bas J Oosterman
Martinus B Oostendorp
Diederick E Grobbee
Kit CB Roes

Therapeutic Innovation and Regulatory Science, 2016; 50(1), 82-90.

Abstract

Background: One approach to increase the efficiency of clinical trial monitoring is to replace 100% Source Data Verification (SDV) by verification of samples of source data. An intuitive strategy for determining appropriate sampling plans (i.e., sample sizes and the maximum tolerable number of transcription errors in the samples) is to use acceptance sampling methodology. Expanding upon earlier work in which the use of acceptance sampling strategies for sampling-based SDV was proposed, we describe an alternative acceptance sampling strategy that, instead of relying on sampling standards, evaluates all possible sampling plans algorithmically, thereby ensuring that selected sampling plans conform to pre-specified criteria. *Methods:* Empirical trial data guided the design of the proposed strategy. In addition, extensive simulations, also based on the empirical data, were performed to assess the performance in terms of workload reductions and the post-SDV error proportion of applying the proposed strategy. *Results:* 13 different scenarios were simulated, but results of the default scenario show that the average pre-SDV error proportion per trial of .056 was reduced to .023 by inspecting only 40.5% of the case report form entries. Of the inspected data entries, almost half (18.0/40.5) was, on average, SDV-ed as part of the sampling process; remaining entries were inspected during full inspections after too many errors were observed in the samples. *Conclusion:* Our results suggest that major reductions in workload can be achieved, while maintaining acceptable data quality levels. However, the results also indicate that the proposed strategy is conservative and that further improvement is possible.

1. Introduction

Source Data Verification (SDV), the process of comparing source data and the data as presented in Case Report Forms (CRFs) during the conduct of clinical trials, serves to ensure completeness, accuracy, reliability and verifiability of data provided by investigators [1]. Despite not being required by regulators, SDV is commonly performed on all CRF entries ('100% SDV'), which is time-consuming, costly, and known to identify errors that have little impact on the trial outcomes [2-8]. Consequently, its use has been a major topic of debate in discussions on the efficiency of clinical trial monitoring, and various alternatives have been proposed [1,4-7,9-15].

One alternative approach (which we term 'random SDV') is to limit SDV to a random sample of CRF entries: Inspection of a sample may be sufficient to assess the quality of all data provided by an investigational site [1,4-7,11]. This view is supported by the Guidelines for Good Clinical Practice, which state that *"Statistically controlled sampling may be an acceptable method for selecting the data to be verified"* [16,p.26] In addition, by reducing the monitors' SDV workload, time becomes available for supporting site staff in increasing the quality.

To determine appropriate sampling plans (i.e. the sample size n and the maximum tolerable number of errors in the sample c), Grieve applied the concept of acceptance sampling to the context of inspection by SDV [11]. Acceptance sampling refers to *"the inspection and classification of a sample of units selected at random from a larger batch or lot and the ultimate decision about the disposition of the lot"* [17,p.15]. If the number of erroneous items in the sample exceeds c, the lot is typically subjected to full inspection. Otherwise, no further inspection is required. For an elaborate introduction to acceptance sampling, see e.g. Montgomery [17].

In the strategy proposed by Grieve [11], the selection of sampling plans relies on a standard sampling scheme in which n and c are provided for a given batch size N and an Acceptable Quality Level (AQL). An advantage of using a sampling standard is that it is practical in implementation. A potential drawback, however, is that the number of possible sampling

plans from which to choose is limited. Therefore, the sampling plans provided by such schemes need not be optimal for the context of the specific inspection (see also Taylor [18] for a discussion on the use of sampling standards).

An alternative approach is to select sampling plans algorithmically, allowing evaluation of every possible sampling plan. Doing so provides flexibility, as the algorithm can be specified to precisely reflect the aims of the inspection procedure, and sampling plans selected by such a procedure are guaranteed to conform to pre-specified criteria.

Here, we describe one possible implementation of the latter approach. We investigate the performance of the proposed strategy versus the 100% SDV approach (in terms of, among others, the workload and effectiveness) using extensive simulations.

The following section provides details concerning the proposed random SDV algorithm. Then, the methodological aspects of the simulation study are provided, followed by its results. We end with a discussion and a conclusion.

2. Proposed acceptance sampling strategy

Grieve [11], in his implementation of random SDV, takes into account that (1) a sampling plan needs to be determined for each specific monitoring visit (MV), and (2) instead of drawing a sample of individual CRF entries (the unit of inspection), it is more feasible to draw samples on the level of patient visits.

Our strategy is in accordance with these considerations. As stated, however, it does not rely on standard sampling schemes, but uses a computational search for the optimal sampling plan. This search requires an estimated value (or range of values) for the expected visit-specific pre-SDV Error Rate (ER). If the true pre-SDV ER of the center is time-dependent (e.g. because of a learning effect), the estimation procedure should take this dependency into account. For this reason, we performed an analysis on empirical trial data to investigate whether such a dependency exists, and concluded it did (see Appendix A for details).

We propose the following algorithm:

1. Right before an SDV visit, randomly order the patient visits entered into the CRF since the previous MV. The possible values for n are then given by the cumulative sum over the number of data entries per patient visit. For each possible value of n, c may equal any value in the set $(0, 1, \dots, n-1)$.
2. Using hypergeometric probabilities, determine β, the probability of accepting the batch if the actual proportion of errors in the batch would be equal to the largest possible proportion of errors smaller than or equal to the Lot Tolerant Percent Defective (LTPD), for every candidate (n, c) pair. Omit the candidate (n, c) pairs for which $\beta > \beta^{spec}$.
3. For the remaining candidate (n, c) pairs, calculate the Average Outgoing Quality Limit (AOQL) as $AOQL = max\left[\frac{P^a \pi (N-n)}{N}\right]$, where π denotes a vector containing every possible error rate (ER), P^a denotes the probability of accepting a lot for every value of π, and N denotes the lot size. Omit any (n, c) pairs for which $AOQL > AOQL^{spec}$.
4. For all remaining (n, c) pairs, calculate the Average Total Inspection (ATI) as $ATI = n + (1 - P^a)(N - n)$.
5. Select the (n, c) pair that minimizes the ATI between π^L and π^U, the values of π representing, respectively, the lowest and upper bound of the expected ER. For the first SDV visit of each site, there is no data yet to base this expectation on, so we set $\pi^L = 0$ and $\pi^U = 1$. For the remaining SDV visits, however, π^L and π^U are determined as follows:
    a. Sample k bootstrap samples from the data on all patient visits inspected thus far.
    b. To each bootstrap sample, fit a logistic regression model, regressing the log odds of observing an error on the time since the first patient visit took place. Use the parameter estimates to predict the number of erroneous entries in each of the conducted, yet uninspected, patient visits that were entered into the CRF since the previous MV. Sum the predicted number of errors over these patient visits and divide this value by N.

c. π^L and π^U are then respectively calculated as the 2.5% and 97.5% percentiles of the resulting distribution of k bootstrap estimates.

As can be seen, the procedure provides sampling plans based on pre-specified values for the LTPD (and corresponding β^{spec}) and $AOQL^{spec}$: Any plan that is returned by the algorithm is guaranteed to have a probability of lot acceptance (i.e., β) at most β^{spec} when the pre-SDV ER is equal to or larger than the LTPD, and the maximum expected long-term post-SDV error rate (the AOQL) will equal at most $AOQL^{spec}$. In addition, it is guaranteed to be the most efficient sampling plan given the predicted pre-SDV ER.

Implementation of the algorithm is preferably achieved through a software environment that is linked to the CRF database. Then, none of the calculations need to be performed by hand. The algorithm only requires the chosen values for the $LTPD$, β^{spec} and $AOQL^{spec}$ as general input at the start of the trial. Through an automated procedure, the system may obtain information on the dates on which each patient visit was entered into the CRF and the number of CRF entries that was or will be required for each patient visit. The only action required from the monitor is to enter the number of errors that he or she detected per inspected patient visit.

3. Methods

3.1. Data simulation

To investigate the performance of the proposed algorithm, we carried out a simulation study in which the algorithm described in the previous section was repeatedly applied to simulated trial data. To simulate trial data, we used the following strategy. For each site s in the trial:

1. Define the recruitment period as the time in week between $t = 0$ and $t = t^{eor}$. Generate the total number of patients included at the site as $N_s^{pat} \sim Pois(\lambda = N^{pat.mean})$. Generate recruitment times for patient $p = 1$ through $p = N_s^{pat}$ as $t_{sp1} \sim Unif(0, t^{eor})$. t^{eor} is set to equal 26 and $N^{pat.mean}$was set to 10. In case $N_s^{pat} = 0$, it was set to equal 1.
2. Set the number of patient visits per patient V equal to 10. Each patient visit v takes place at time $t_{spv} = t_{sp1} + v \cdot \Delta t^{pv}$, where Δt^{pv}, the number of weeks between two

consecutive patient visits, was set to 13. For patient visit v, ranging from 1 to V, the corresponding number of CRF entries is indicated by D_v. We set D_v equal to (80, 25, 25, …, 25, 50).

3. For every time point t_{spv} on which a patient visit takes place, calculate the probability of entering data into the CRF incorrectly P^e_{spv} as the inverse logit of $\beta_0^* + \xi_s^* + (\beta_1^* + \eta_s^*)t_{spv}$, where β_0^* and β_1^* were set to equal $\hat{\beta}_0$ and $\hat{\beta}_1$, and ξ_s^* and η_s^* are simulated as $(\xi_s^*, \eta_s^*) \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\hat{\sigma}_\xi^2 & \hat{\sigma}_{\xi,\eta}\\ \hat{\sigma}_{\xi,\eta} & \hat{\sigma}_\eta^2\end{pmatrix}\right]$, as estimated in Appendix A.
4. Then, simulate the actual number of erroneous CRF entries per patient visit E_{spv} as $E_{spv} \sim Binom(D_v, P^e_{spv})$. Use the data generated thus far to set up patient visit level data matrix $\boldsymbol{M}_s^{pv}$ with number of rows equal to $V \cdot N_s^{pat}$, each row representing a patient visit.
5. The number of weeks between consecutive MVs on which SDV was performed Δt^m was set to 26. The time of each MV m was determined as $min(t_{spv}) + m \cdot \Delta t^m$, except for the final MV, which was set to $max(t_{spv}) + 0.5$.
6. In $\boldsymbol{M}_s^{pv}$, include the value for m representing the earliest MV after the patient visit.
7. Lastly, generate an MV-level data matrix $\boldsymbol{M}_s^m$ with each row representing an MV and columns specifying s, m, the number of patient visits per MV, the number of CRF entries per MV and the number of erroneous CRF entries per MV.

Trial-level data was obtained by combining $S = 100$ site-level data sets.

3.2. Choice of $LTPD$, β^{spec} and $AOQL^{spec}$

Implementation of the proposed acceptance sampling algorithm requires specification of $LTPD$, β^{spec} and $AOQL^{spec}$. These values require careful consideration, but it must be noted that, because (1) the AOQL itself is a limit, (2) we select (n, c)-pairs based on two criteria, (3) n is always equal to or larger than strictly required due to sampling patient visits instead of individual CRF entries, and (4) P^a and π are discrete for finite N, the results are likely to be conservative. Therefore, the $LTPD$, β^{spec} and $AOQL^{spec}$ should not be chosen too strict. In our simulations, we used $LTPD = 0.1$, $\beta^{spec} = 0.1$ and $AOQL^{spec} = 0.05$. In practice, the

selection of appropriate values for these criteria is highly trial-specific, e.g. depending on whether the data is to be used in the analysis and whether the trial is blinded.

3.3. Inspection error

Typically, the inspection process is not perfect [17]. In his simulations, Grieve [11] assumed that 95 percent of the errors will be detected and corrected when using 100% SDV, but indicated that this rate might be too optimistic. In our simulations, we assumed a detection rate of 90%.

3.4. Responsibility of full inspection after lot rejection

We assumed that it is the monitor who performs the full inspection after a lot has been rejected. One might, however, argue that these full inspections are the responsibility of the site staff. In that case, it is of relevance to assess the distribution of workload between the monitor and the site. Our results allow such assessments, if it is assumed that the monitor and the site staff are equally effective in detecting and correcting errors, and that no re-inspection is performed on data already fully inspected by the site staff.

3.5. Outcomes

Five trial-level outcomes were considered: The pre-SDV ER indicates the ER that would be observed if no inspection would have been employed at all. The 'post-SDV ER (100% SDV)' denotes the ER after the data has been subjected to 100% SDV. The 'post-SDV ER (random SDV)' equals the ER after the proposed random SDV strategy has been applied to the data. The proportion SDV-ed (total) indicates the total workload, i.e. the total proportion of CRF entries that required SDV. The fifth trial-level outcome is the proportion SDV-ed (in-sample), denoting the proportion of CRF entries that required SDV, including only the CRF entries that were part of the initial samples (i.e., excluding CRF entries that were SDV-ed as part of the post-rejection full inspections).

In addition, batch sizes, absolute and relative sample sizes, the probability of lot rejection, and the values of β and AOQL of the selected sampling plans were investigated, albeit not on the trial level, but on the MV level, combining all MVs from all trials simulated in each setting.

3.6. Simulation setup

In the data simulation algorithm, the following parameters were varied:

1. The number of sites per trial S, set to 100 in the default setting, was increased and decreased to 200 and 50.
2. The follow-up duration in terms of V, the number of patient visits required. In the default setting, $V = 10$, and was increeased and decreased to 20 and 5.
3. The time between two consecutive MVs Δt^m, set to 26 in the default setting, was increased and decreased to 52 and 13.
4. The pre-SDV ER, in terms of fixed intercept β_0^*, set to $\hat{\beta}_0$ in the default setting, was set to $\hat{\beta}_0 + \log(0.5)$ and $\hat{\beta}_0 + \log(2)$, resulting in overall pre-SDV error rates that are 0.5 and 2 times the overall pre-SDV error rate as estimated from the data (see Appendix A).

Besides varying parameters in the data simulation algorithm, we varied the LTPD and $AOQL^{spec}$. The LTPD equals .1 in the default setting, and was in- and decreased to 0.2 and 0.05. $AOQL^{spec}$, set to 0.05 in the default setting, was in- and decreased to 0.1 and 0.025.

The parameters were varied univariately. Per setting, 100 trials were simulated. For settings in which no adjustment in the data simulation algorithm was required (i.e. the settings in which LTPD and $AOQL^{spec}$ were varied), the data simulated for the default setting were used. For practical reasons, the number of bootstrap samples k was limited to 500.

4. Results

4.1. Trial-level outcomes

Table 1 and Figure 1 show the results of the simulation study, in terms of the distributions of the trial-level outcomes in each simulated setting. In the default setting, the average ER per trial if no SDV would have been performed equals 0.056. If SDV was performed using 100% SDV, this ER decreases to 0.006. If, instead, the proposed random SDV strategy was used, the average post-SDV ER equals 0.023, and it would require, on average, SDV of 40.5 percent of the CRF entries, almost half (18.0/40.5) of which was SDV-ed as part of the sampling process (the remaining half was inspected after lot rejections).

Setting	Pre-SDV ER	Post-SDV ER (100% SVD)	Post-SDV ER (random SVD)	Prop. SDV-ed (total)	Prop. SDV-ed (in-sample)
Default	0.056 (0.004)	0.006 (0.0004)	0.023 (0.0010)	0.405 (0.024)	0.180 (0.004)
50 sites	0.057 (0.006)	0.006 (0.0007)	0.023 (0.0015)	0.409 (0.038)	0.181 (0.008)
200 sites	0.057 (0.003)	0.006 (0.0003)	0.023 (0.0007)	0.407 (0.016)	0.180 (0.004)
5 patient visits	0.072 (0.004)	0.007 (0.0005)	0.028 (0.0014)	0.491 (0.032)	0.188 (0.005)
20 patient visits	0.046 (0.006)	0.005 (0.0006)	0.017 (0.0009)	0.321 (0.024)	0.149 (0.005)
MV every 52 wks	0.056 (0.004)	0.006 (0.0004)	0.026 (0.0014)	0.348 (0.029)	0.144 (0.004)
MV every 13 wks	0.056 (0.004)	0.006 (0.0004)	0.020 (0.0008)	0.476 (0.021)	0.232 (0.005)
Low pre-SDV ER	0.030 (0.003)	0.003 (0.0003)	0.019 (0.0009)	0.251 (0.019)	0.163 (0.004)
High pre-SDV ER	0.104 (0.006)	0.010 (0.0007)	0.024 (0.0011)	0.612 (0.026)	0.191 (0.005)
LTPD = 0.05	*See default*	0.006 (0.0004)	0.011 (0.0006)	0.673 (0.022)	0.224 (0.005)
LTPD = 0.2	*See default*	0.006 (0.0004)	0.031 (0.0014)	0.276 (0.023)	0.121 (0.008)
$AOQL^{spec}$= 0.025	*See default*	0.006 (0.0004)	0.018 (0.0009)	0.484 (0.024)	0.150 (0.008)
$AOQL^{spec}$= 0.1	*See default*	0.006 (0.0004)	0.023 (0.0010)	0.402 (0.023)	0.180 (0.004)

Table 1. Means (standard deviations) of the trial-level outcomes in each of the simulated settings. All estimates are based on 100 simulated trials.

As can be seen, the average outcomes are not structurally affected by the number of sites per the trial. There is, however, a clear inverse association between the number of sites and the variability of the outcomes, as was expected.

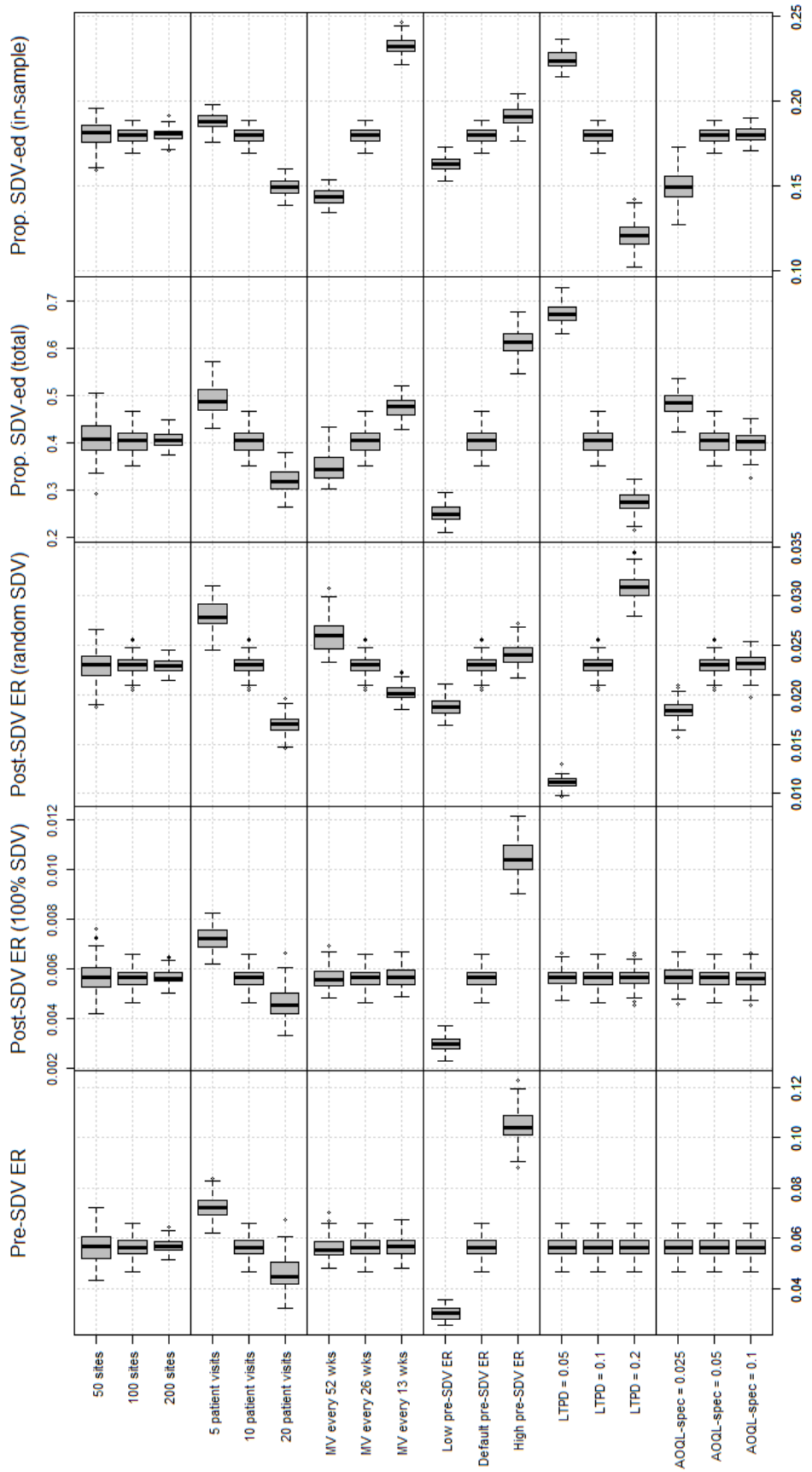

Figure 1. Distributions of the trial-level outcomes in each of the simulated settings. Each boxplot is based on 100 simulated trials.

Increasing the duration of the trial (by increasing the number of required patient visits) decreases the pre-SDV ER. The post-SDV ER for 100% SDV and the workload for the random SDV strategy decrease as a direct consequence. It can also be seen that the post-SDV ER for the random SDV strategy decreases. Presumably this finding is due to the increased ability to make predictions: not only is less effort required, the effort that is required is used more effectively.

Varying the time between two consecutive MVs does not affect the pre-SDV ER or the post-SDV ER for 100% SDV. This is not surprising, since our simulations assume that there is no relation between the findings of SDV and future ERs, and the detection rate is assumed independent of N. However, there appears to be an effect on both the proportion SDV-ed and the post-SDV ER for random SDV. With larger lot sizes, fewer data entries require inspection. In this case, however, we do see an increase in the average post-SDV ER.

Lower pre-SDV ERs result in a lower post-SDV ER, for both 100% SDV and random SDV, as compared to the default. In addition, there is a clear effect of the pre-SDV ER on the workload for the random SDV strategy: in the 'high pre-SDV ER' setting, 61.2% of the CRF entries required SDV, compared to 25.1% in the 'low pre-SDV ER' setting. As can be seen, this large difference is only to a small extent caused by differences in the relative sizes of the samples; it is mainly the result of the increased number of required post-rejection full inspections.

The effects of varying the LTPD and $AOQL^{spec}$ are as anticipated: if set stricter, lower post-SDV ERs are observed, at the cost of increased SDV workload. This finding appears to hold true especially for the LTPD. Note, however, that there appears to be no structural difference between the results of the default scenario and the scenario in which $AOQL^{spec}$= 0.1, due to the AOQLs of the sampling plans generated by the algorithm having an upper bound determined by β^{spec} (and vice versa). An additional interesting observation is that decreasing $AOQL^{spec}$ results in a higher overall workload, but the proportion of that workload spend on inspection of samples is reduced.

4.2. MV-level outcomes

Combining all MVs from all trials simulated in the default setting, N (the number of CRF entries since the previous MV), equals, on average, 554.3. The sample size n, on the other hand, equals 99.6 on average. The average relative sample size n/N equals 0.21. Overall, in around 25% of the samples, the number of detected erroneous CRF entries exceeds the critical value c, resulting in full inspection of the N CRF entries. On average, sampling plans generated by the algorithm correspond to a β of .08 (close to β^{spec} = .10) and an AOQL of 0.028 (smaller than $AOQL^{spec}$ = .05). These results can be found in Table 2.

Setting	N	n	n/N	Proportion rejected	Observed β	Observed AOQL
Default	554.3 (289.8)	99.6 (56.1)	0.21 (0.13)	0.25	0.080 (0.022)	0.028 (0.009)
50 sites	551.2 (289.8)	99.4 (55.8)	0.21 (0.14)	0.25	0.080 (0.022)	0.028 (0.009)
200 sites	554.7 (290.5)	99.7 (56.1)	0.21 (0.13)	0.25	0.080 (0.022)	0.028 (0.009)
5 patient visits	683.4 (295.4)	128.4 (57.7)	0.20 (0.08)	0.36	0.086 (0.014)	0.032 (0.007)
20 patient visits	529.9 (242.1)	79.1 (53.2)	0.17 (0.12)	0.19	0.075 (0.025)	0.024 (0.009)
MV every 52 wks	1097.4 (433.2)	157.6 (84.3)	0.15 (0.07)	0.22	0.083 (0.020)	0.035 (0.009)
MV every 13 wks	302.6 (146.6)	70.2 (37.6)	0.25 (0.12)	0.31	0.079 (0.018)	0.023 (0.006)
Low pre-SDV ER	553.0 (289.4)	89.9 (51.0)	0.19 (0.13)	0.11	0.078 (0.024)	0.027 (0.009)
High pre-SDV ER	553.0 (291.1)	105.6 (65.7)	0.22 (0.15)	0.47	0.080 (0.021)	0.027 (0.009)
LTPD = 0.05	*See default*	124.2 (66.8)	0.26 (0.14)	0.53	0.081 (0.019)	0.010 (0.004)
LTPD = 0.2	*See default*	67.0 (70.5)	0.16 (0.17)	0.14	0.018 (0.021)	0.043 (0.009)
$AOQL^{spec}$= 0.025	*See default*	83.0 (83.9)	0.18 (0.17)	0.33	0.058 (0.031)	0.021 (0.004)
$AOQL^{spec}$= 0.1	*See default*	99.7 (56.2)	0.21 (0.13)	0.25	0.080 (0.023)	0.028 (0.009)

Table 2. The MV-level outcomes in each of the simulated settings. For each setting, all MVs from all trials were combined. For all outcomes except the 'proportion rejected', means are provided (with standard deviations between brackets).

In table 2, the MV-level outcomes of the other settings can be found as well. Since these outcomes are no longer on the trial level, they are unaffected by the number of sites per trial. However, when the follow-up duration (in terms of the number of patient visits) is increased, both N and n decrease. The reason for this finding is that the average number of CRF entries per patient visit becomes smaller. There appears to be no clear effect on n/N, but the probability of batch rejection decreases as a result of the effect of time on the pre-SDV ERs. In addition, the β and AOQL values of the sampling plans generated by the algorithm appear to decrease as the number of patient visits increases. Likely, this observation is a result of P^a and π becoming more discrete for smaller values of N.

Increasing the time between consecutive MVs results, by definition, in an increase in N. It also increases n, but not proportionally, as the difference in n/N shows. Also, the probability of lot rejection decreases. Probably, this finding is related to the effect of the ER over time; the first MVs will likely result in rejection of the batch, unless the MV is performed late enough to ensure that the quality of the inspected batch is sufficient again. The β and AOQL values of the sampling plans generated by the algorithm increase. The latter observation is, again, likely the result of the increased lot sizes.

Increasing the pre-SDV error proportions appears to primarily affect the probability of rejecting the batch, as can be expected. In the low pre-SDV ER setting, the decision to perform full inspection was made in 11 percent of the SDV visits, versus 47 percent in the high pre-SDV ER setting. In addition, there might be a small effect of the pre-SDV ER on the selected sample sizes, but differences are small compared to the overall variation in the simulation results.

Increasing the LTPD results in a decrease of n (and, as a consequence, of n/N). In addition, fewer lots are rejected. In the results for the observed β and AOQL values, the dependency between the two becomes apparent: For the low and default values of the LTPD, the AOQL values are conservative, as they are bounded by β^{spec}. For the LTPD = 0.2 setting, in contrast, the observed average value of β (0.018) is conservative, while the AOQL is, on average, closer to $AOQL^{spec}$.

Similarly, increasing the $AOQL^{spec}$ beyond its default value of 0.05 does not affect the results. Decreasing its value, however, results in a larger rejection probability (as can be expected) and in generated sampling plans that are conservative with respect to β, but not with respect to the AOQL.

Note that an overview of the main characteristics of the simulated data is presented in Appendix B.

5. Discussion

These results suggest significant cost savings may result from using acceptance sampling in the context of SDV: Although dependent on specific trial characteristics, post-SDV error rates range from 1 to 3.5 percent in the investigated scenarios. The proportion SDV-ed typically ranges from .3 to .6 (unless the pre-SDV ER is very high or low relative to the pre-specified criteria).

To appreciate these findings, some aspects of the study need to be addressed: (1) Acceptance sampling is a reactive approach, and unlikely to improve quality levels of future batches. It is therefore advisable to implement acceptance sampling in combination with other methodologies aimed to improve the quality. (2) There is a large variation in workload, even within settings, making it difficult to determine/estimate the required workload (and the monitoring budget) a priori. (3) The acceptance sampling strategy assumes a single definition of what constitutes an error. This implies that between-monitor variation in error definitions should be reduced as much as possible, by providing a clear and generally applicable error definition. Since such a definition was not present in the trial from which the empirical data was obtained, we simply considered any entries that were later revised as being erroneous. This might have affected our empirical analysis (and as a consequence, our simulation study).

There are certain limitations to the proposed acceptance sampling algorithm. Sampling entire patient visits (instead of individual CRF entries) may be practical, but, generally speaking, reduces the representativeness of the samples. Also, the results indicate that the strategy is conservative, implying that values for the LTPD and $AOQL^{spec}$ should not be chosen too strict, which makes it difficult to determine the optimal values. Another potential problem is that missing CRF fields may not be detected as such. For example, if a lab result did not reach the physician, the CRF entry form will not be generated, so SDV will not detect this value as being missing. This issue might be dealt with by considering SDV and checks for completeness as two separate processes. Also, it should be noted that the algorithm proposed here serves as an example or starting point. As the aims of inspection may differ per trial, the algorithm may require adjustment according to specific trial characteristics or aims. For instance, one might want to include the AQL after all (to limit the chances of incorrect lot

rejections, at the cost of larger sample sizes), make β^{spec} a function of N, consider double sampling plans, etc. Another possible adjustment would be to take inspection error into account formally (see, e.g. Govindaraju [19]).

Further limitations are related to the set-up of the simulation study. For example, drop-outs or deaths were not simulated, a fixed inspection detection rate of .9 was assumed, the frequency of MVs was independent of the number of patients included per site, etc. In practice, the simulation algorithm may need to be adapted to more closely resemble the specific trial characteristics. Note also that the simulation setup differs to a large extent from the setup that was chosen by Grieve [11], and drawing comparisons in terms of the results should be done with caution.

We limited our focus to the sampling procedure itself. We did not discuss the effect of errors on the final analysis. Although the observed post-inspection ERs appear sufficiently low for many purposes (i.e. below 3.5 percent), we did not investigate whether such error rates have the potential to affect a trial's results. It would be interesting to perform further simulation studies on this topic as well (see also Grieve [11] for a discussion on this topic).

Finally, random SDV is not the only alternative to 100% SDV. Other approaches that have been proposed are, e.g., 'targeted SDV' (i.e., focus SDV on high risk data points), 'remote SDV' (allowing SDV to be performed centrally), or hybrid approaches [1,4-7,9-15]. The TransCelerate Biopharma Inc. guidance on risk-based monitoring estimated that, of all generated queries, the percentage generated through SDV is as low as 7.8% overall and 2.4% in critical data [20]. These numbers may suggest not performing SDV at all. In addition, strategies to limit the possibility for transcription errors to occur in the first place (e.g., reducing complexity of CRFs or the amount of data collected, improving automated data entry checks, improved CRF training, direct data entry, etc.) should be considered.

6. Conclusion

We proposed a random SDV strategy based on acceptance sampling theory, similar to Grieve [11], but one that is more flexible, allowing for adjustments to guarantee correspondence

between the selected sampling plans and the actual aims of SDV inspection. We showed that material efficiency gains may be achieved when replacing a 100% source data verification approach with an acceptance sampling approach, particularly in trials with lower pre-SDV ERs and/or long follow-up.

As a next step, we recommend that this acceptance sampling approach is implemented in an actual trial, to determine its characteristics and practical feasibility more realistically. Also, it is interesting to investigate which post-SDV ER levels are actually acceptable for various types of trials and analyses. In addition, it is recommended that regulatory agencies be more specific concerning their views on sampling-based approaches to SDV.

Appendix A: Analysis of empirical trial data

The data on which this analysis was performed came from a multi-center, multinational phase III trial and were obtained from the trial's data mutation logs. These logs contain information on every instance that a data field was entered into the CRF or that an existing entry was altered. For each of over 700 sites, we extracted the weekly number of initial CRF entries, and determined whether or not these entries were later modified. If so, the initial CRF entry was considered erroneous, unless the latest modification was performed within two days after the initial entry, as those cases are unlikely to have been encountered by a monitor during a site visit. We excluded CRF entries for which the initial entry was either (1) not performed by the site, or (2) not part of a patient visit after randomization. Due to specifics of the mutation logging system, a small portion (0.15 percent of all entries in the final dataset) of initial entry dates were intractable (while the corresponding mutation dates were not). These dates were therefore singly imputed using a random draw from the observed distribution of days between initial entries and modifications of that site.

To these data, a logistic mixed effects regression model was fitted, regressing the weekly proportion of erroneous CRF items (for weeks in which at least one CRF data field was submitted) on the number of weeks since the first CRF entry. More specifically, we fitted the model

$$logit(Y) = \beta_0 + \xi_s + (\beta_1 + \eta_s)t,$$

where Y denotes the binary variable indicating whether a CRF entry was erroneous or not, β_0 denotes the fixed intercept, ξ_s denotes the random intercept for each site, and β_1 and η_s denote, respectively, the fixed and random effect of t, the number of weeks since the first CRF entry. An effect of time was assumed if β_1 could be shown to be significantly different from zero (i.e. if its p-value < 0.05).

Using the glmer function in the lme4 package for R [21,22], we estimated $\hat{\beta}_0 = -2.235$ (SE = 0.026) and $\hat{\beta}_1 = -0.017$ (SE = 0.0005, Z = -33.70, Wald p-value < 2e-16). Therefore, we assumed the presence of a time effect. The random effect parameters were estimated as $\hat{\sigma}_\xi^2 = 0.4709$, $\hat{\sigma}_\eta^2 = 0.0002$ and $\hat{\sigma}_{\xi,\eta} = -0.0041$.

Figure A1 shows, for all sites, the site-specific predicted proportion of erroneous CRF entries as a function of the number of weeks since first CRF entry. Furthermore, the overall curve based on the fixed effect estimates is displayed. As can be seen, the overall predicted proportion of erroneous CF entries equals 0.096 at $t = 0$, and decreases to 0.011 at $t = 130.5$, the median follow-up time.

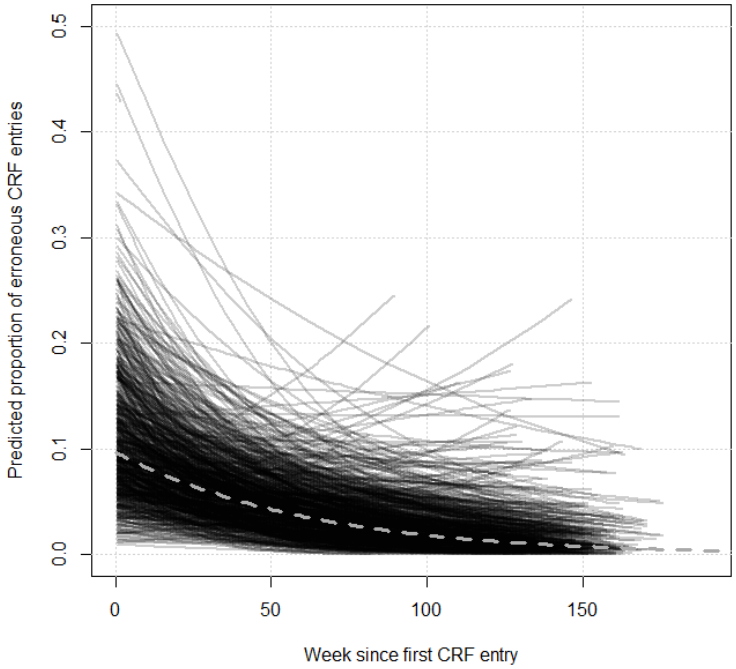

Figure A1: Predicted proportions of erroneous CRF entries as a function of the number of weeks since first CRF entry per site (black curves) and overall (dashed grey curve).

Appendix B: Simulated Trial Data

Table B1 provides an overview of important characteristics of the simulated trial data. For example, in the default setting, $S = 100$, so 100 sites were simulated per trial. As $N^{pat.mean} = 10$, the average number of patients included per trial approximately equals 1000 (here, 1001.3). Because $V = 10$, the average number of patient visits is 10 times 1001.3. Also, since $\Sigma D_v = 330$, the average number of CRF entries per trial equals 330 times 1001.3. The number of MVs per site will typically equal 6, but may sometimes equal 5 if the distribution of t_{sp1} on a specific site happens to contain low values only. Per trial, the average number of MVs should therefore be close to 100 times 6, which corresponds to the observations.

Setting	Number of sites	Number of patients	Number of patient visits	Number of CRF entries	Number of MVs
Default	100	1001.3 (32.4)	10012.7 (324.4)	330419.1 (10706.0)	596.1 (1.9)
50 sites	50	497.5 (22.7)	4975.4 (227.2)	164188.2 (7498.9)	297.9 (1.4)
200 sites	200	2003.7 (44.2)	20037.4 (442.0)	661234.2 (14587.2)	1192.1 (2.8)
5 patient visits	100	1000.1 (34.8)	5000.4 (173.8)	205018.5 (7125.8)	300
20 patient visits	100	1001.5 (30.0)	20030.6 (600.6)	580887.4 (17418.0)	1096.1 (1.9)
MV every 52 wks	100	997.6 (27.9)	9976.1 (279.1)	329211.3 (9211.0)	300
MV every 13 wks	100	1004.8 (33.6)	10047.6 (336.4)	331570.8 (11102.3)	1095.9 (2.1)
Low pre-SDV ER	100	998.6 (26.9)	9986.3 (269.2)	329547.9 (8883.8)	595.9 (1.8)
High pre-SDV ER	100	998.2 (29.4)	9981.6 (294.3)	329392.8 (9711.3)	595.7 (2.2)

Table B1. Means (standard deviations) of basic trial-level characteristics of the simulated trial data. All estimates are based on 100 simulated trials. Note that the settings consisting of variations of the LTPD or $AOQL^{spec}$ use the data from the default setting, and are therefore not separately provided in the table.

References

1. Khosla R, Verma DD, Kapur A, Khosla S. Efficient source data verification. Indian J of Pharmacol. 2000; 32:180-186.
2. Eisenstein EL, Lemons PW II, Tardiff BE, Schulman KA, King Jolly M, Califf RM. Reducing the costs of phase III cardiovascular clinical trials. Am Heart J. Mar 2005; 149(3):482-488.
3. Funning S, Grahnén A, Eriksson K, Kettis-Linblad Å. Quality assurance within the scope of good clinical practice (GCP) - what is the cost of GCP-related activities? A survey within the Swedish association of the pharmaceutical industry (LIF)'s members. Qual Assur J. 2009; 12:3-7.
4. Tantsyura V, Grimes I, Mitchel J, et al. Risk-based source data verifications: pros and cons. Drug Inf J. 2010; 44:745-756.
5. De S. Hybrid approaches to clinical trial monitoring: practical alternatives to 100% source data verification. Perspect Clin Res. 2011; 2(3):100-104.
6. Getz K. Low hanging fruit in the fight against inefficiency. Appl Clin Trials. Mar 2011.
7. Hines S. Targeting source document verification. Applied Clinical Trials. Feb 2011.
8. Tudor Smith C, Stocken DD, Dunn J, et al. The value of source data verification in a cancer clinical trial. PLOS ONE. 2012; 7(12):e51623.
9. Radovich C, Frick J. Remote source document verification (rSDV): a sponsor perspective and results of implementation. The Monitor. Dec 2009.
10. Bakobaki JM, Rauchenberger M, Joffe N, McCormack S, Stenning S, Meredith S. The potential for central monitoring techniques to replace on-site monitoring: findings from an international multi-centre clinical trial. Clin Trials. 2012; 9:257-264.
11. Grieve AP. Source data verification by statistical sampling: issues in implementation. Drug Inf J. 2012; 46:368-377. 2012.
12. Venet D, Doffagne E, Burzykowski T, et al. A statistical approach to central monitoring of data quality in clinical trials. Clin Trials. 2012; 9(6):705-713.
13. FDA. Guidance for industry oversight of clinical investigations: a risk-based approach to monitoring. 2013.

14. Nielsen E, Hyder D, Deng C. A Data-Driven Approach to Risk-Based Source Data Verification. Ther Innov Regul Sci. 2014; 48(2):173-180.
15. Uren SC, Kirkman MB, Dalton BS, Zalcberg JR. Reducing clinical trial monitoring resource allocation and costs through remote access to electronic medical records. J Onc Prac. 2013; 9(1):13-15.
16. ICH. ICH harmonized tripartite guideline- guideline for good clinical practice E6(R1). 1996.
17. Montgomery DC. Introduction to statistical quality control. John Wiley & Sons, Inc. 2009. ISBN 9780470169926.
18. Taylor WA. Selecting statistically valid sampling plans. Qual Eng 1997; 10(2):365-370.
19. Govindaraju K. Inspection error adjustment in the design of single sampling attributes plan. Qual Eng 2007; 19:227-233.
20. TransCelerate BioPharma Inc. Position paper: Risk-based monitoring methodology. 2013. URL: http://www.transceleratebiopharmainc.com/wp-content/uploads/2013/10/TransCelerate-RBM-Position-Paper-FINAL-30MAY2013.pdf
21. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. URL: http://www.R-project.org/.
22. Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. 2014. URL: http://CRAN.R-project.org/package=lme4.

Chapter 3

A computationally simple central monitoring procedure, effectively applied to empirical trial data with known fraud

Rutger M van den Bor
Petrus WJ Vaessen
Bas J Oosterman
Nicolaas PA Zuithoff
Diederick E Grobbee
Kit CB Roes

Journal of Clinical Epidemiology, 2017; 87, 59-69.

Abstract

Objectives: Central monitoring of multi-center clinical trials becomes an ever more feasible quality assurance tool, in particular for the detection of data fabrication. More widespread application, across both industry-sponsored as well as academic clinical trials, requires central monitoring methodologies that are both effective and relatively simple in implementation. *Study design and setting:* We describe a computationally simple fraud detection procedure intended to be applied repeatedly and (semi-)automatically to accumulating baseline data and to detect data fabrication in multi-center trials as early as possible. The procedure is based on anticipated characteristics of fabricated data. It consists of seven analyses, each of which flags approximately 10% of the centers. Centers that are flagged three or more times are considered 'potentially fraudulent' and require additional investigation. The procedure is illustrated using empirical trial data with known fraud. *Results:* In the illustration data, the fraudulent center is detected in the majority of repeated applications to the accumulating trial data, while keeping the proportion of false positive results at sufficiently low levels. *Conclusion:* The proposed procedure is computationally simple and appears to be effective in detecting center-level data fabrication. However, assessment of the procedure on independent trial datasets with known data fabrication is required.

1. Introduction

The exact prevalence of fraud or data fabrication[1] in clinical trials is difficult to estimate, but generally assumed to be low [1-6]. Whether it has a substantial impact on a trial's outcomes depends on the extent and nature of the fraudulent behavior [1,5,7]. However, *"even isolated and small amounts of fraud within a trial can cause significant doubts about its conclusions and have the potential to lead to a lack of public confidence for the clinical trial process in general"* [5,p.226]. Moreover, being aware of recent cases of fraud may deter individuals from participating in clinical research [8]. Therefore, we agree with Friedman et al. [9,p.42], who state: *"We condemn all data fabrication. It is important to emphasize that confidence in the integrity of the trial and its results is essential to every trial. If, through intentional or inadvert actions, that confidence is impaired, not only have the participants and potentially others in the community been harmed, the trial loses its rationale, which is to influence science and medical practice"*.

Traditionally, quality of trial data is monitored by on-site monitoring and source data verification. However, various authors (e.g. [1,4,5,7,10-12]) have argued that central site-by-site comparisons of digitally available clinical trial data may be more effective than on-site monitoring visits in detecting fraud. For examples of central fraud detection strategies, see [1,2,4,5,11,13-17].

Fraud detection strategies commonly rely on statistical significance tests that are aimed to assess whether a center-specific data pattern deviates (in terms of e.g. means, variances, digit preference, etc.) from the overall data pattern. The decision to flag a center is then based on the resulting p-value. Although this general strategy may yield informative results, the assumption that low p-values are indicative for relevant deviations can be problematic, as (1) substantial structural variability is often observed between centers [11] (making a null hypothesis of no difference unrealistic to start with) and (2) the number of observations (i.e. recruited subjects) per center often is highly variable. If so, centers with relatively many observations will be structurally disadvantaged, as was observed by Kirkwood et al. [4, p.789]: *"When a large number of data values were examined, even a small difference […]*

[1]We use the terms 'fraud' and 'data fabrication' interchangeably.

sometimes produced a small p-value […]". In addition, if the central monitoring procedure is to be performed on an ongoing basis and/or the number of statistical tests is large, it easily becomes infeasible to assess whether the assumptions of the significance tests are met.

In this paper, we propose an alternative strategy to detect possible fraud in multi-center trials. The strategy shares the aim of detecting deviating data patterns on the center level, but uses a weighting procedure, rather than significance tests, to take into account differences in center-specific recruitment numbers. We illustrate the performance of the strategy by applying it repeatedly to accumulating empirical baseline data from a trial with known fraud.

2. Proposed fraud detection strategy

The proposed strategy consists of seven analyses that are based on anticipated characteristics of fabricated data, each of which returns a selection of approximately 10% of the centers that are 'most suspicious'. The total number of analyses on which a center is flagged then serves as the basis for determining whether a center requires closer inspection. Centers are included if they recruited a minimum of five subjects. All analyses were programmed and performed in R [18]. Details are provided in the following sections.

2.1. Anticipated characteristics of fabricated data

A fraudulent staff member typically does not have access to any trial data besides the data from the subjects recruited by the specific center for which the staff member works. Consequently, we anticipate that, for continuously measured variables, distributions of fabricated observations will be different from the true observations. Fabricated data values may be, on average, too low or too high [1,4,5,19]. Also, data fabrication may become apparent by investigating the spread of the distributions [1,4,5,19,20]. Specifically, we expect variability to be lower in fabricated data, because fraudulent investigators either choose to refrain from fabricating extreme values to avoid triggering attention or simply underestimate the variability [20]. Bivariately, deviations may become apparent when comparing pair-wise correlations [1,2,4,5,20]. These expectations are assessed in analyses 1, 2 and 3.

In some respects, fabricated data may be expected to be 'too perfect'. We anticipate that fabricated data will contain relatively few missing values. Also, we expect that the rate by which patients are recruited will be relatively constant over time, as a result of inclusion of either ineligible patients and/or phantom subjects. Analysis 4 and 5 concern these expectations.

Another potential indication of fraud, assessed in analysis 6, is based on the notion that fraudulent investigators may fail to take into account the relative irregularity of subject visits taking place during weekends [1,4,5,7]. Finally, in analysis 7, we compared the distribution of first, second, or last digits [1,2,4,5,7,19]. This analysis is similar to the commonly employed fraud detection procedure based on Benford's law [21], but using empirical, rather than assumed, reference distributions, as in [4].

2.2. Indicators

The next step concerns the quantification of the expectations outlined in the previous section. This quantification should take into account the possibility that the data are not yet extensively cleaned [11], and should therefore be robust against outliers.

We use the following indicators: For analysis 1, the 'average difference' is quantified using the 'common language effect size' of the Wilcoxon-Mann-Whitney test, comparing each center-specific distribution against the distribution observed when combining the patient data from all other centers. This measure varies from 0 to 1, and is interpreted as the proportion of possible pairs on which the value belonging to sample 1 is larger than the value belonging to sample 2 [22,23]. For analysis 2, the difference in spread is quantified as the natural logarithm of the ratio of inter-quartile ranges. The difference in pair-wise correlations (analysis 3) is quantified as the estimated absolute difference in Kendall's τ correlations. For analysis 4, we use the difference in the proportion of missing data entries. The stability by which subjects are recruited (analysis 5) is quantified as the average difference between the observed cumulative proportion of recruited subjects as a function of time and the linear slope indicating a perfectly stable recruitment pattern. Values close to 0 indicate a stable recruitment pattern, while values close to 1 indicate non-stability. For analysis 6, we use the difference in

proportions of visits taking place during the weekend. Finally, to compare distributions of digits in analysis 7, we use the dissimilarity index, interpreted as the proportion of observations that needs to be adjusted in order to make the digit distribution on the center of interest equal to the digit distribution based on the observations from all other centers.[2]

2.3. Accounting for variation in the number of observations per center

Each center's score on each of the indicators can be estimated from the observed data. However, when variation exists in center-specific recruitment numbers, a direct comparison of centers based on these estimates is problematic, because the estimates can be expected to be more variable when based on few observations. That is, it is not surprising to observe relatively extreme scores for centers with few recruited subjects. We account for this variation by means of a computationally simple and pragmatic weighting procedure that shrinks the estimated indicator values towards a certain null value indicative of 'no data fabrication'. Denote the number of subjects recruited by center i as n_i. Then, the weighted estimates are equal to the sum of $\frac{n_i}{m+ n_i}$ times the estimated indicator value and $\frac{m}{m+ n_i}$ times the null value. Constant m represents a tuning parameter that determines the rate by which the weighted estimate converges from the null value to the estimate itself with increasing number of observations, and can informally be interpreted as a pseudo-count. Due to the difficulty in determining an optimal value for m a priori, we repeat the analyses using multiple values (0, 5, 10 and 20). Note that, in case an analysis is repeated for multiple variables, we average the variable-specific scores. For further details, we refer the reader to the R programs available from the Web Appendix[3].

2.4. Selection of potentially fraudulent centers

In each analysis, we flag the subset of centers with indicator values that, after weighting, correspond to the 10% most extreme values. Therefore, each center can be flagged up to

[2]Note that the dissimilarity index is known to suffer from an upward bias when dealing with few observations and/or a population value close to 0 [24]. For this reason, we chose to apply the bias correction procedure proposed by Mazza and Punzo [24] (specifically, we used the estimator denoted $\widetilde{D}_{Boot}$ in the original article).

[3]See http://www.jclinepi.com/article/S0895-4356(16)30398-5/abstract

seven times. Assuming that the chance of being flagged in each of the seven analyses indeed equals 10%, and assuming those probabilities are independent, we expect approximately half (48%) of the centers will not be flagged at all, 85% at most once, and 97% at most twice. Therefore, it appears reasonable to consider centers that are flagged three or more times to be worthy of increased inspection.

3. Illustration on empirical trial data

3.1. The Second European Stroke Prevention Study (ESPS2)

The randomized, placebo-controlled, double-blind ESPS2 trial aimed to investigate the safety and efficacy of low-dose acetylsalicylic acid, modified-release dipyridamole, and a combination of the two agents for the prevention of ischemic stroke. The trial population consisted of patients who were at least 18 years old and who had experienced a transient ischemic attack or a completed ischemic stroke within the preceding three months. The analysis was based on data from 6602 patients that were recruited in 59 centers. Subjects were followed for two years. Further details and results are provided elsewhere [25].

However, data submitted by an additional center with 438 subjects had been excluded due to suspected fraud. In this case, initial doubts arose during standard monitoring, after which an investigation was carried out by the sponsor. This investigation revealed various atypical data patterns, such as unusually high recruitment rates and very low adverse event rates.

The identity of 97% of the patients was retrieved, a subset of which was studied in more detail. This investigation revealed that, while the majority of patients did experience a qualifying event, there was no indication in the records that follow-up visits had taken place for the purpose of the trial, and patients (or their general practitioners) were unaware of their involvement in the trial [11,26,27]. For a detailed account of the discovery and handling of this specific case of fraud, see [26] and/or [27].

For the purpose of designing a fraud detection algorithm, we were given access to the baseline Case Report Form (CRF) data of the trial. Excluding conditional items (e.g. ‘if cardiac failure is present, specify the stage according to the NYHA classification’), the CRF consisted of 161

items, of which 18 were (semi-)continuous measurements (e.g. height, weight, sedimentation rate), 139 were binary or categorical measurements (e.g. smoking status, type of cerebrovascular accident, normal versus abnormal coordination of limbs), and 5 variables indicated dates (e.g. date of birth, date of cerebrovascular accident, date of planned next visit).

Note, however, that the fraudulent investigator may not have truly fabricated all baseline CRF entries. As stated, most patients did experience a qualifying event, and information obtained as part of routine care may have been used without consent. However, we anticipate that fabrication would have been necessary for at least part of the CRF.

The center-specific cumulative recruitment patterns as a function of the days since the first randomization in the trial are provided in Figure 1. Only few centers recruited more than 438 patients.

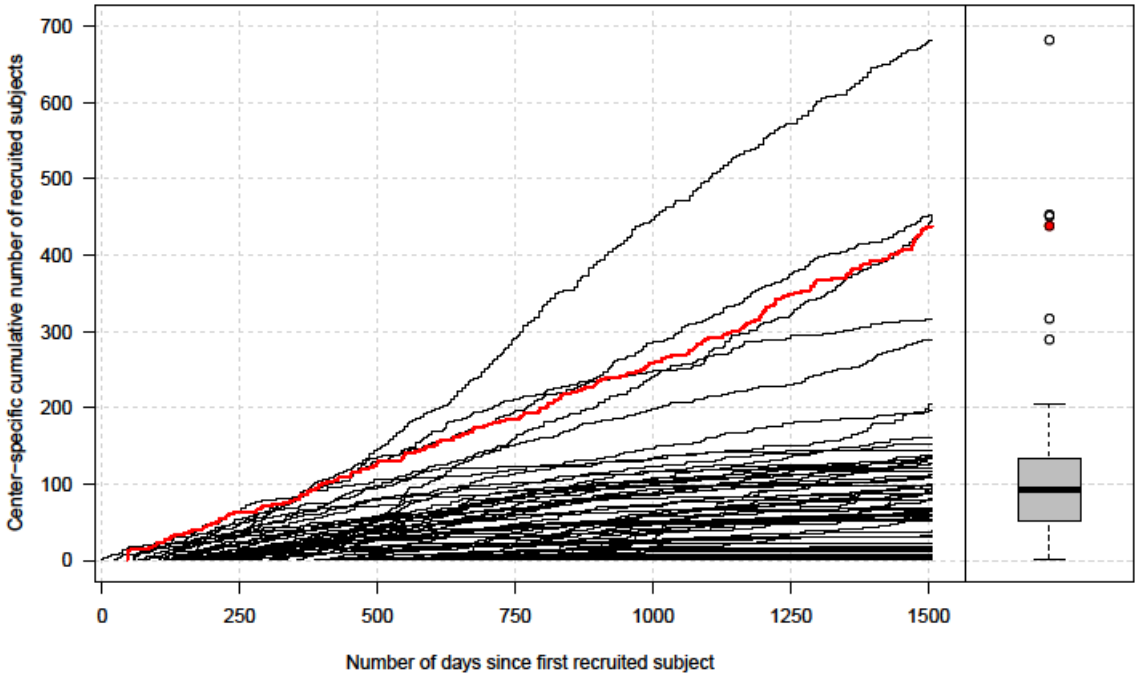

Figure 1. Overview of cumulative recruitment patterns as a function days since the first subject was recruited, for all 60 centers in the ESPS2 trial. The boxplot on the right shows the distribution of total center-specific subject numbers. The fraudulent center is displayed in red.

3.2. Methods

3.2.1. Data used in the analyses

Analysis 1, 2 and 3 all use the set of continuous variables (height, weight, systolic blood pressure, diastolic blood pressure, heart rate during general examination, heart rate during ECG, leucocytes, erythrocytes, platelets, haematocrit, haemoglobin, erythrocyte sedimentation rate, urea, creatinine, uric acid, fasting glucose, cholesterol, cholesterol LDL), as well as the set of variables indicating date differences (age, days since CVA, days since ECG measurements, days until the next subject visit). Analysis 4 uses the overall proportion of missing values, and is therefore based on all 161 items. Analysis 5 compares recruitment rates, thus requiring only the date of the baseline visit. Analysis 6 compares the proportion of subject visit dates taking place during weekends. Here, three such dates are available and used: the date of the ECG, the date of the baseline visit, and the planned date of the next visit. Analysis 7 is based on the same items as analyses 1, 2 and 3, but using the date values directly (i.e. the date of birth, the date of the CVA, the date of the ECG, the date of the baseline visit, and the planned date of the next visit).[4] Note that, with very few observations, computational problems quickly become unavoidable. Therefore, centers with fewer than 5 subjects will not be considered for potential detection, and variables with very few observations are automatically removed from the analysis (again, see the R-programs in the Web Appendix for details).

3.2.2. Investigated scenarios

The fraudulent center in the ESPS2 trial (‘center 2013’) was initiated relatively early. In addition, its investigator ‘recruited’ a relatively large number of subjects (438 in total). Consequently, the number of subjects recruited in center 2013 (denoted by n_{2013}) is relatively high at any moment during the recruitment period. Yet, it is of interest to assess the ability of the algorithms to detect the data fabrication in other, perhaps less optimal, scenarios as well.

[4]For the majority of variables we focus on the second digit, because variation in the leading digit typically is limited. Only for diastolic blood pressure do we use the leading digit, because the measurements are frequently rounded to multiples of five. For date variables, we use the day of the month and focus on the last digit.

Therefore, we altered the data of center 2013 in various ways, yielding twelve different scenarios. An overview is presented in Table 1.

	Early vs. late initiation of center 2013		
	Early	*Late (a)*	*Late (b)*
Recruitment not halted	1 (438)*	1a (150)	1b (194)
Recruitment halted at median	2 (92)	2a (92)	2b (92)
Recruitment halted at 25th percentile	3 (52)	3a (52)	3b (52)
Recruitment halted at 10th percentile	4 (15)	4a (15)	4b (15)

*Table 1. Overview of the scenarios used in the analysis. The numbers in brackets indicate the total number of subjects recruited in center 2013 in each scenario. *Setting 1 represents the empirical scenario.*

In scenario 1, the data is used as it was observed empirically. In scenarios 2, 3 and 4, we decrease n_{2013}: In scenario 2, we halt the recruitment after 92 subjects (the median number of recruited subjects per center in the trial) are included. In scenario 3 and 4, we limit the number of subjects even further, to 52 and 15, respectively (the 25th and 10th percentile number of recruited subjects per center).

In these scenarios, however, center 2013 is still one of the first centers to be initiated. Therefore, we manipulate the data in order to resemble the situation in which center 2013 was initiated last. We do so in two ways: First, we 'shift' recruitment (and all other) dates by 130 weeks (scenarios 1a, 2a, 3a and 4a). Second, subjects recruited by center 2013 before the last center was initiated are removed from the data (scenarios 1b, 2b, 3b and 4b). Note that, because the end of the recruitment period is left unadjusted, these manipulations do affect the total number of subjects recruited by center 2013 for scenarios 1a and 1b.

3.2.3. Repeated application to accumulating trial data

The fraud detection strategy described in this paper is intended to be computationally simple, so that it can easily be applied repeatedly to accumulating trial baseline data and to detect potential cases of fraud (if it occurs) as early as possible.

We apply the data fabrication detection procedure to the accumulating empirical baseline data every fourth week, but only after at least 5 centers are initiated. Therefore, the first run takes

place on the first day on which at least 5 centers recruited at least 5 subjects. For these data, this moment takes place $T = 59$ days after the first subject was recruited. The last two runs (run number 52 and 53) are performed at $T = 1487$ and $T = 1507$, the latter being the day on which the last subject was recruited. Note that, since center 2013 was one of the first centers to be initiated, run 1 (at $T = 59$) is not performed in the scenarios where the initiation date is delayed.

3.3. Results

As indicated, a center is considered potentially fraudulent when, in a given run of the fraud detection strategy, it is flagged on at least three of the seven analyses. In Table 2, we show the proportion of runs on which the fraudulent center is flagged zero through seven times, for each investigated scenario and every m. E.g. in simulation scenario 1 with $m = 10$, center 2013 was flagged two times in 4 percent of the runs, three times in 53 percent of the runs and four times in 43 percent of the runs. These results confirm the feasibility of using a threshold of being flagged on at least three out of seven analyses, as it appears to be the largest value that still results in frequent detection of center 2013 in most scenarios.

Scenario	*m*	Number of times flagged								Flagged 3 or more times
		0	1	2	3	4	5	6	7	
1*	0	0.13	0.13	0.15	0.53	0.06	-	-	-	0.58
	5	-	0.02	0.09	0.64	0.25	-	-	-	0.89
	10	-	-	0.04	0.53	0.43	-	-	-	0.96
	20	-	-	-	0.13	0.75	0.09	0.02	-	1.00
1a	0	-	-	-	0.80	0.20	-	-	-	1.00
	5	-	-	-	0.80	0.20	-	-	-	1.00
	10	-	-	-	0.85	0.15	-	-	-	1.00
	20	-	-	0.05	0.50	0.45	-	-	-	0.95
1b	0	-	-	0.25	0.50	0.25	-	-	-	0.75
	5	-	-	0.30	0.30	0.40	-	-	-	0.70
	10	-	-	0.35	0.25	0.40	-	-	-	0.65
	20	0.05	-	0.30	0.30	0.35	-	-	-	0.65
2	0	0.13	0.04	0.38	0.45	-	-	-	-	0.45
	5	-	0.02	0.11	0.85	0.02	-	-	-	0.87
	10	-	-	0.04	0.91	0.06	-	-	-	0.96
	20	-	-	-	0.87	0.09	0.02	0.02	-	1.00
2a	0	-	-	-	1.00	-	-	-	-	1.00
	5	-	-	-	1.00	-	-	-	-	1.00
	10	-	-	-	1.00	-	-	-	-	1.00
	20	-	-	0.05	0.70	0.25	-	-	-	0.95
2b	0	-	-	0.80	0.15	0.05	-	-	-	0.20
	5	-	-	0.85	0.10	0.05	-	-	-	0.15
	10	-	-	0.90	0.10	-	-	-	-	0.10
	20	0.05	-	0.85	0.10	-	-	-	-	0.10
3	0	0.13	0.06	0.19	0.62	-	-	-	-	0.62
	5	-	0.02	0.08	0.89	0.02	-	-	-	0.91
	10	-	-	0.04	0.87	0.09	-	-	-	0.96
	20	-	-	-	0.04	0.92	-	0.04	-	1.00
3a	0	-	-	-	1.00	-	-	-	-	1.00
	5	-	-	-	1.00	-	-	-	-	1.00
	10	-	-	-	1.00	-	-	-	-	1.00
	20	-	-	0.05	0.10	0.85	-	-	-	0.95
3b	0	-	-	0.20	0.75	0.05	-	-	-	0.80
	5	-	-	0.50	0.45	0.05	-	-	-	0.50
	10	-	-	0.65	0.35	-	-	-	-	0.35
	20	0.05	-	0.85	0.10	-	-	-	-	0.10
4	0	0.04	0.02	0.09	0.83	0.02	-	-	-	0.85
	5	-	-	0.13	0.85	0.02	-	-	-	0.87
	10	-	-	0.30	0.68	0.02	-	-	-	0.70
	20	-	-	0.96	0.02	0.02	-	-	-	0.04
4a	0	-	-	-	1.00	-	-	-	-	1.00
	5	-	-	-	1.00	-	-	-	-	1.00
	10	-	-	-	1.00	-	-	-	-	1.00
	20	-	-	1.00	-	-	-	-	-	0.00
4b	0	-	-	-	-	1.00	-	-	-	1.00
	5	-	-	-	0.15	0.85	-	-	-	1.00
	10	-	-	0.05	0.60	0.35	-	-	-	0.95
	20	0.05	-	0.90	0.05	-	-	-	-	0.05

*Table 2. For each investigated scenario and for varying m, the proportion of runs on which the fraudulent center is flagged 0 through 7 times, and the proportion of runs on which the fraudulent center was flagged at least three times. Values equal to 0 are represented by '-'. *Scenario 1 represents the empirical scenario.*

Table 3 shows, for each simulation scenario, (1) the proportion of runs on which the fraudulent center was flagged on each of the analyses, and (2) the proportion of runs on which the fraudulent center was flagged at least three times.

	Early initiation of center 2013				*Late initiation of center 2013*							
	*Scenario 1 ($n_{2013} = 438$)**				*Scenario 1a ($n_{2013} = 150$)*				*Scenario 1b ($n_{2013} = 194$)*			
	$m = 0$	$m = 5$	$m = 10$	$m = 20$	$m = 0$	$m = 5$	$m = 10$	$m = 20$	$m = 0$	$m = 5$	$m = 10$	$m = 20$
Analysis 1 ('average' values)	0.00 (0)	0.00 (0)	0.02 (0)	0.45 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Analysis 2 (variability)	0.60 (1)	0.94 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)	0.95 (1)
Analysis 3 (pair-wise correlations)	0.00 (0)	0.00 (0)	0.00 (0)	0.04 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.25 (0)	0.05 (0)	0.05 (0)	0.05 (0)	0.00 (0)
Analysis 4 (missing values)	0.21 (0)	0.28 (0)	0.30 (0)	0.34 (0)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Analysis 5 (recruitment pattern)	0.60 (0)	0.64 (1)	0.66 (1)	0.68 (1)	0.20 (1)	0.20 (1)	0.15 (1)	0.20 (1)	0.45 (1)	0.50 (1)	0.50 (1)	0.45 (1)
Analysis 6 (visits during weekends)	0.08 (0)	0.25 (0)	0.42 (1)	0.49 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.50 (0)	0.55 (1)	0.55 (1)	0.55 (1)
Analysis 7 (digit preference)	0.75 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)
Flagged 3 or more times	0.58 (0)	0.89 (1)	0.96 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)	0.75 (1)	0.70 (1)	0.65 (1)	0.65 (1)
	Scenario 2($n_{2013} = 92$)				*Scenario 2a ($n_{2013} = 92$)*				*Scenario 2b ($n_{2013} = 92$)*			
	$m = 0$	$m = 5$	$m = 10$	$m = 20$	$m = 0$	$m = 5$	$m = 10$	$m = 20$	$m = 0$	$m = 5$	$m = 10$	$m = 20$
Analysis 1 ('average' values)	0.00 (0)	0.00 (0)	0.02 (0)	0.09 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Analysis 2 (variability)	0.45 (1)	0.91 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)	0.95 (1)
Analysis 3 (pair-wise correlations)	0.00 (0)	0.00 (0)	0.00 (0)	0.04 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.25 (0)	0.05 (0)	0.05 (0)	0.05 (0)	0.00 (0)
Analysis 4 (missing values)	0.87 (1)	0.94 (1)	0.96 (1)	0.98 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Analysis 5 (recruitment pattern)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.10 (0)	0.10 (0)	0.10 (0)	0.10 (0)
Analysis 6 (visits during weekends)	0.02 (0)	0.02 (0)	0.04 (0)	0.08 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.10 (0)	0.05 (0)	0.00 (0)	0.00 (0)
Analysis 7 (digit preference)	0.81 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)
Flagged 3 or more times	0.45 (1)	0.87 (1)	0.96 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)	0.20 (0)	0.15 (0)	0.10 (0)	0.10 (0)
	Scenario 3 ($n_{2013} = 52$)				*Scenario 3a ($n_{2013} = 52$)*				*Scenario 3b ($n_{2013} = 52$)*			
	$m = 0$	$m = 5$	$m = 10$	$m = 20$	$m = 0$	$m = 5$	$m = 10$	$m = 20$	$m = 0$	$m = 5$	$m = 10$	$m = 20$
Analysis 1 ('average' values)	0.00 (0)	0.00 (0)	0.04 (0)	0.13 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Analysis 2 (variability)	0.62 (1)	0.96 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)	0.95 (1)
Analysis 3 (pair-wise correlations)	0.00 (0)	0.00 (0)	0.00 (0)	0.87 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.85 (1)	0.05 (0)	0.05 (0)	0.05 (0)	0.00 (0)
Analysis 4 (missing values)	0.81 (1)	0.91 (1)	0.94 (1)	0.96 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Analysis 5 (recruitment pattern)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.70 (1)	0.45 (0)	0.35 (0)	0.10 (0)
Analysis 6 (visits during weekends)	0.02 (0)	0.04 (0)	0.08 (0)	0.08 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.10 (0)	0.05 (0)	0.00 (0)	0.00 (0)
Analysis 7 (digit preference)	0.85 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)
Flagged 3 or more times	0.62 (1)	0.91 (1)	0.96 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)	0.80 (1)	0.50 (0)	0.35 (0)	0.10 (0)
	Scenario 4 ($n_{2013} = 15$)				*Scenario 4a ($n_{2013} = 15$)*				*Scenario 4b ($n_{2013} = 15$)*			
	$m = 0$	$m = 5$	$m = 10$	$m = 20$	$m = 0$	$m = 5$	$m = 10$	$m = 20$	$m = 0$	$m = 5$	$m = 10$	$m = 20$
Analysis 1 ('average' values)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Analysis 2 (variability)	0.92 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)	0.95 (1)
Analysis 3 (pair-wise correlations)	0.02 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (1)	0.85 (1)	1.00 (1)	0.05 (1)
Analysis 4 (missing values)	0.91 (1)	0.87 (1)	0.70 (1)	0.04 (0)	1.00 (1)	1.00 (1)	1.00 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Analysis 5 (recruitment pattern)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Analysis 6 (visits during weekends)	0.02 (0)	0.02 (0)	0.02 (0)	0.02 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (1)	1.00 (1)	0.35 (0)	0.00 (0)
Analysis 7 (digit preference)	0.91 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.95 (1)
Flagged 3 or more times	0.85 (1)	0.87 (1)	0.70 (1)	0.04 (0)	1.00 (1)	1.00 (1)	1.00 (1)	0.00 (0)	1.00 (1)	1.00 (1)	0.95 (1)	0.05 (1)

*Table 3. Per simulation scenario, the proportion of runs on which the fraudulent center was flagged (per analysis), and the proportion of runs on which the fraudulent center was flagged at least three times. The values of 0 and 1 between brackets indicate whether center 2013 was flagged in the last run (performed at $T = 1507$ days after the first subject was included in the trial).*Scenario 1 represents the empirical scenario.*

For scenario 1, it can be observed that the overall probability of detection is high (at least 0.58) and increases to unity with increasing m. The effectiveness appears to be primarily driven by analysis 2 and 7, both of which flag center 2013 frequently, most prominently when m equals 5, 10 or 20. The fraudulent center was also commonly flagged in analysis 4, 5 and 6. In contrast, analysis 1 and 3 appear relatively ineffective.

When reducing the number of subjects recruited in center 2013 (scenarios 2, 3 and 4) or when postponing the initiation of center 2013 by shifting (scenarios 1a, 2a, 3a, and 4a), the

probability of detecting center 2013 remains high, the only exception being scenario 4 and 4a with $m = 20$. For analyses 1, 2, 3 and 7, performance is comparable to the performance in scenario 1. However, the effectiveness of analysis 5 and 6 is reduced, while the performance of analysis 4 improves.

When the initiated date of center 2013 is postponed by removing the subjects recruited first (scenarios 1b, 2b, 3b, and 4b), results are more ambiguous. The probability of detection is drastically reduced in scenario 2b and 3b. Overall, the proportion of runs on which the fraudulent center is flagged is high for analysis 2 and 7 and low for analysis 1 and 4. The performance of the remaining analyses, however, varies considerably per scenario.

Table 4 provides information regarding the probability with which the procedure falsely identifies non-fraudulent centers as being potentially fraudulent: (1) the median proportion of false positive detections over the repeated runs, (2) the false positive rate on the last run, and (3) the unique number of centers detected at least once over the full set of runs (note that the total number of centers equals 60). No material differences are observed between the different scenarios. Two observations can be made: First, the false positive rate is relatively high (around 0.09) with $m = 0$, but decreases to around 0.02 with increasing m. Second, the number of centers that is detected at least once, while of course dependent on the number of runs, can be limited by choosing a sufficiently large value for m.

Scenario	m	Median (Q1, Q3) false positive rate	False positive rate on last run	Unique false positive detections (count)
1*	0	0.09 (0.07, 0.10)	0.09	31
	5	0.04 (0.03, 0.05)	0.04	16
	10	0.04 (0.02, 0.05)	0.02	13
	20	0.02 (0.02, 0.04)	0.02	7
1a	0	0.09 (0.07, 0.09)	0.09	30
	5	0.04 (0.03, 0.06)	0.04	19
	10	0.04 (0.02, 0.05)	0.02	16
	20	0.02 (0.02, 0.04)	0.02	14
1b	0	0.09 (0.07, 0.09)	0.09	30
	5	0.04 (0.03, 0.06)	0.04	21
	10	0.04 (0.02, 0.05)	0.02	14
	20	0.02 (0.02, 0.04)	0.02	14
2	0	0.09 (0.07, 0.09)	0.09	31
	5	0.04 (0.03, 0.05)	0.04	17
	10	0.04 (0.02, 0.05)	0.02	12
	20	0.02 (0.02, 0.04)	0.02	9
2a	0	0.09 (0.07, 0.09)	0.09	30
	5	0.04 (0.04, 0.06)	0.02	20
	10	0.04 (0.02, 0.05)	0.02	15
	20	0.02 (0.02, 0.04)	0.02	14
2b	0	0.09 (0.07, 0.09)	0.09	30
	5	0.04 (0.02, 0.06)	0.02	20
	10	0.04 (0.02, 0.05)	0.02	15
	20	0.02 (0.02, 0.04)	0.02	14
3	0	0.09 (0.07, 0.09)	0.09	31
	5	0.04 (0.02, 0.05)	0.02	16
	10	0.04 (0.02, 0.05)	0.02	12
	20	0.02 (0.02, 0.04)	0.02	9
3a	0	0.09 (0.07, 0.09)	0.09	30
	5	0.04 (0.03, 0.05)	0.02	21
	10	0.04 (0.02, 0.05)	0.02	14
	20	0.02 (0.02, 0.04)	0.02	14
3b	0	0.09 (0.07, 0.09)	0.09	30
	5	0.04 (0.03, 0.05)	0.04	22
	10	0.04 (0.02, 0.05)	0.02	14
	20	0.02 (0.02, 0.04)	0.02	14
4	0	0.08 (0.06, 0.09)	0.09	27
	5	0.04 (0.02, 0.05)	0.02	18
	10	0.04 (0.02, 0.05)	0.02	11
	20	0.02 (0.02, 0.04)	0.02	10
4a	0	0.09 (0.07, 0.09)	0.09	30
	5	0.04 (0.04, 0.06)	0.04	20
	10	0.04 (0.02, 0.05)	0.02	15
	20	0.02 (0.02, 0.04)	0.02	14
4b	0	0.07 (0.07, 0.09)	0.07	30
	5	0.04 (0.02, 0.05)	0.02	20
	10	0.04 (0.02, 0.05)	0.02	14
	20	0.03 (0.02, 0.04)	0.02	14

*Table 4. For every scenario and m, the median (and Q1 and Q3) false positive rate over the runs, the false positive rate on the last run, and the unique number of non-fraudulent centers that were detected on at least one of the runs (note: in total, 60 centers were initiated. On the last run, the analysis included 57 non-fraudulent centers, as two centers recruited fewer than 5 subjects). *Scenario 1 represents the empirical scenario.*

With respect to the choice of m, choosing a larger value reduces the false positive rate, but complicates the detection of a fraudulent center (if present) if that center recruited a limited number of subjects. Our results suggest that using $m = 5$ or $m = 10$ yields a reasonable trade-off.

For the empirical scenario, detailed graphical results are provided in Figure 2. In it, each center is represented by a horizontal line. For each of the 53 runs of the detection procedure, the color of the line segment indicates the number of analyses on which the center was flagged. These results show that, besides center 2013, one other center (the center marked with an 'X') is consistently flagged three or more times. This center, however, included only five patients in total. Although additional monitoring would have been advisable had the trial still been ongoing, we expect this finding to be a false positive result.

4. Discussion

In this paper, we described a computationally simple procedure for the detection of data fabrication in multicenter clinical trials. The procedure was illustrated retrospectively on empirical trial data with known fraud. When applied repeatedly to the accumulating trial data, the procedure detects the fraudulent center in the majority of runs. Letting pseudo-count m equal 5 or 10, detection rates amount to 87 or 96 percent. In the empirical data, the fraudulent center was initiated early in the recruitment period and 'recruited' a relatively large number of subjects. However, with few exceptions, good performance was maintained when postponing the center's initiation date or limiting its 'recruitment'. The proportion of false positive findings on any run was typically low, with median proportions never exceeding 0.04 when m equals 5, 10, or 20. Of the seven analyses that constitute the detection procedure, the analysis comparing the variability of continuous distributions and the analysis concerning digit preference, i.e. analyses 2 and 7, appeared to be most informative in the empirical data. Least informative were the analyses comparing the location of continuous distributions (analysis 1) and the analysis comparing pair-wise correlations (analysis 3). It should be determined whether these results are generalizable to other cases of data fabrication.

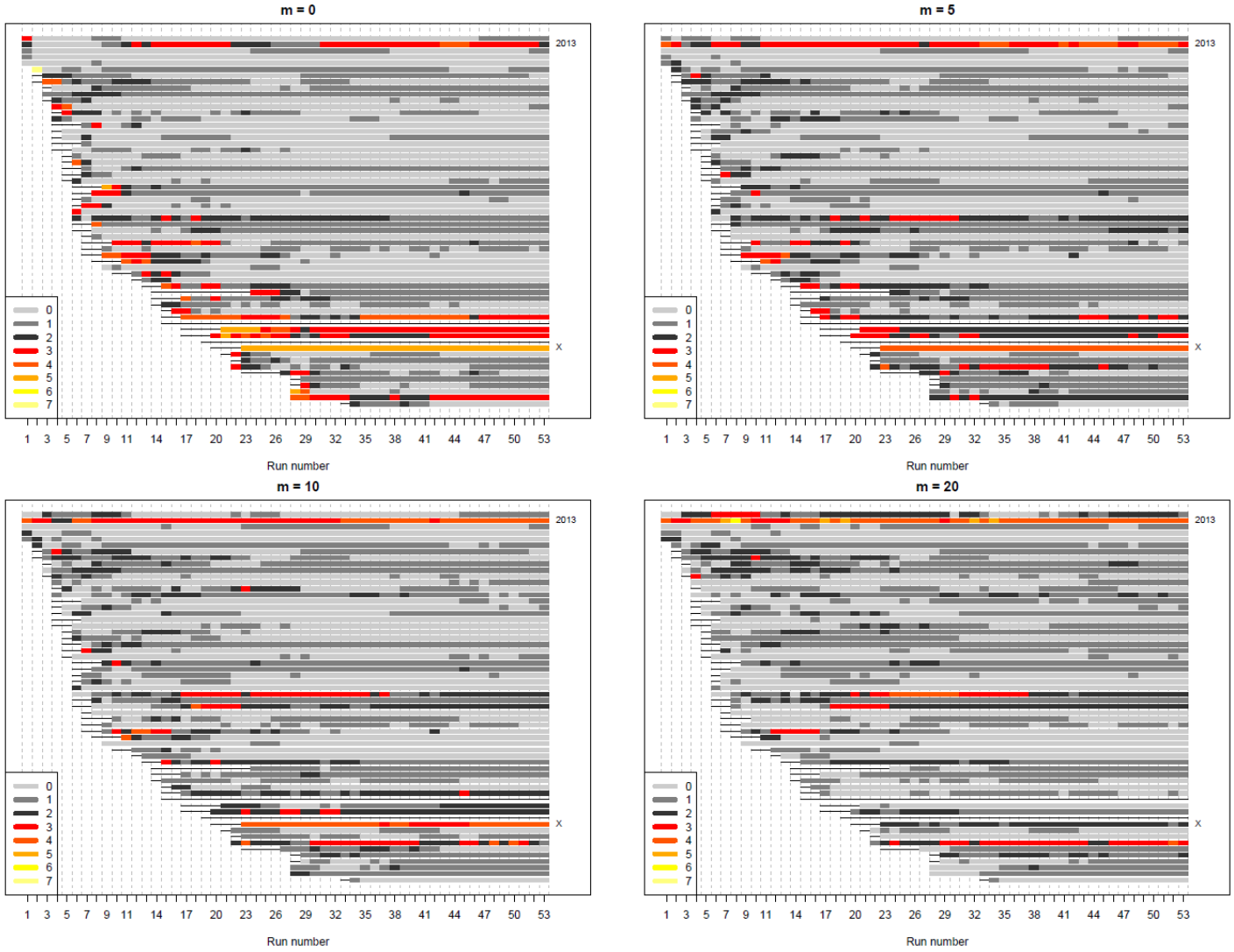

Figure 2. Graphical results of repeatedly applying the data fabrication procedure to the accumulating ESPS2 trial data for scenario 1, with m equal to 0, 5, 10 and 20. Each horizontal line represents a center. Colors represent the number of analyses on which the center was flagged. The second line corresponds to center 2013, i.e. the center where fraud was detected. Narrow black lines indicate that the center did start recruiting subjects at the time of the run, but was not included because the number of recruited subjects was lower than 5.

There are several potential limitations related to the proposed fraud detection strategy. First, it is based on assumptions regarding the characteristics of fabricated data, and these assumptions need not always hold. Second, out of the seven analyses, four utilize variables that are measured on a (semi-)continuous scale. Therefore, it can be problematic to apply the procedure to a trial in which no or only few of such variables are collected. Third, the procedure is specifically intended to discover data fabrication and will therefore not necessarily aid in detecting other types of errors, although similar procedures could perhaps be constructed for other purposes as well. Finally, a problem with the repeated application of the procedure to accumulating data is that, while the proportion of false positive findings may be limited when assessed for each run separately, the total number of centers that would require additional investigation over the course of recruitment period as a whole may still be relatively large. Although this number can be reduced by increasing m or by reducing the number of runs, a more stringent selection strategy may need to be considered.

Improvements to the proposed strategy may be possible. First, the current strategy uses baseline data only. Although this keeps the strategy simple in implementation and enables its application early in the trial, additional analyses that make use of repeated measures or adverse event data may be incorporated. Second, we chose m to be constant. In practice, one might consider using analysis- or even variable-specific values for m.

Although the procedure appears promising when applied to the ESPS2 data, we stress that the application serves as an illustration, and not as a formal performance analysis, which would require (preferably blinded) assessment using independent trial data. Also, the two data manipulation approaches aimed to postpone the initiation of the fraudulent center have their drawbacks: The first approach ignores possible changes in the trial protocol or seasonal effects, and the second approach treats subjects that were actually recruited relatively late as if they were among the first subjects to be recruited by the center.

Quantitative assessment of the performance of fraud detection strategies is difficult due to the limited availability of empirical trial data containing known data fabrication. As a consequence, comparing the performance of the detection procedure proposed here to other

strategies is difficult as well, although doing so might be an interesting focus of future research. We invite those with access to data with known fraudulent cases to independently validate the proposed approach and provide suggestions for adaptation or further refinement. The R programs that were used are available from the Web Appendix.

5. Conclusion

Central monitoring of multi-center clinical trials becomes an ever more feasible quality assurance tool, and may be particularly useful for the detection of fraud. Common strategies to detect fraud rely on comparisons of center-specific P-values, which might be problematic when the number of subjects recruited per center varies, and which are not always robust against data entry errors. We propose an alternative, computationally simple fraud detection procedure, and have illustrated its application on empirical trial data with known fraud. In these data, the probability of detecting the fraudulent center was, in general, high, while false positive rates were low.

References

1. Buyse M, George SL, Evans S, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. Stat Med 1999; 18(24):3435-3451.
2. Taylor RN, McEntegart DJ, Stillman EC (2002). Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. Ther Innov Regul Sci 2002; 36(1):115-125. DOI: 10.1177/009286150203600115.
3. Al-Marzouki S, Roberts I, Marshall T, et al. The effect of scientific misconduct on the results of clinical trials: A Delphi survey. Contemp Clin Trials 2005;26(3):331-337.
4. Kirkwood AA, Cox T, Hackshaw A. Application of methods for central statistical monitoring in clinical trials. Clin Trials 2013; 10(5):783-806. DOI: 10.1177/1740774513494504.
5. Pogue JM, Devereaux PJ, Thorlund K, et al. Central statistical monitoring: Detecting fraud in clinical trials. Clin Trials 2013; 10(2):225-235. DOI: 10.1177/1740774512469312.
6. George SL. Research misconduct and data fraud in clinical trials: prevalence and causal factors. Int J Clin Oncol 2016; 21:15-21. DOI: 10.1007/s10147-015-0887-3
7. Baigent C, Harrell FE, Buyse M, et al. Ensuring trial validity by data quality assurance and diversification of monitoring methods. Clin Trials 2008; 5(1):49-55. DOI: 10.1177/1740774507087554.
8. Garimella P, Martin IG. Influence of clinical research investigator fraud on clinical trial participation. Ther Innov Regul Sci 2013; 47(1):90-94. DOI: 10.1177/0092861512457776.
9. Friedman LM, Furberg CD, DeMets DL, et al. Fundamentals of clinical trials. 5th ed. Cham Heidelberg New York Dordrecht London: Springer, 2015.
10. Usher RW. PhRMA BioResearch monitoring committee perspective on acceptable approaches for clinical trial monitoring. Ther Innov Regul Sci 2010; 44(4):477-483. DOI: 10.1177/009286151004400412.
11. Venet D, Doffagne E, Burzykowski T, et al. A statistical approach to central monitoring of data quality in clinical trials. Clin Trials 2012; 9(6):705-713. DOI: 10.1177/1740774512447898.

12. Food and Drug Administration. Guidance for industry: Oversight of clinical investigations - A risk-based approach to monitoring. August 2013.
13. Wu X, Carlsson M. Detecting data fabrication in clinical trials from cluster analysis perspective. Pharmaceutical statistics 2011; 10:257-264. DOI: 10.1002/pst.462
14. Desmet L, Venet D, Doffagne E, et al. Linear mixed-effects models for central statistical monitoring of multicenter clinical trials. Stat Med 2014; 33(30):5265-5279. DOI: 10.1002/sim.6294.
15. Timmermans C, Venet D, Burzykowski T. Data-driven risk identification in phase III clinical trials using central statistical monitoring. Int J Clin Oncol. Epub ahead of print 2 aug 2015. DOI: 10.1007/s10147-015-0877-5.
16. 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(2):144-154. DOI: 10.1177/2168479016630576
17. Oba K. Statistical challenges for central monitoring in clinical trials: A review. Int J Clin Oncol 2016; 21:28–37. DOI: 10.1007/s10147-015-0914-4
18. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. URL: http://www.R-project.org/.
19. Al-Marzouki S, Evans S, Marshall T, et al. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. BMJ 2005; 331:267-270. DOI: http://dx.doi.org/10.1136/bmj.331.7511.267.
20. Knatterud GL, Rockhold FW, George SL, et al. Guidelines for quality assurance in multicenter trials: A position paper. Control Clin Trials 1998; 19(5):477-493.
21. Benford F. The law of anomalous numbers. Proceedings of the American Philosophical Society 1938; 78(4):551-572.
22. Conroy RM. What hypotheses do "nonparametric" two-group tests actually test? Stata J 2012; 12(2):182-190.
23. Kerby DS. The simple difference formula: an approach to teaching nonparametric correlation. Innovative Teaching 2014; 3(1):Article 1. DOI: 10.2466/11.IT.3.1.
24. Mazza A, Punzo A. On the upward bias of the dissimilarity index and its corrections. Socio Meth Res 2015; 44(1):80-107. DOI: 10.1177/0049124114543242.

25. Diener HC, Cunha L, Forbes C, et al. European Stroke Prevention Study 2. Dipyridamole and acetylsalicylic acid in the secondary prevention of stroke Neurol Sci 1996; 143(1-2):1-13.
26. Hoeksema HL, Troost J, Grobbee DE, et al. Fraud in a pharmaceutical trial. Lancet 2000; 356(9243):1773.
27. Hoeksema HL, Troost J, Grobbee DE, et al. Een geval van fraude bij farmaceutisch onderzoek in de neurologie [A case of fraud in a neurological pharmaceutical clinical trial]. Ned Tijdschr Geneeskd 2003; 147(28):1372-1377.

Chapter 4

Predicting enrollment performance of investigational centers in phase III multi-center clinical trials

Rutger M van den Bor
Diederick E Grobbee
Bas J Oosterman
Petrus WJ Vaessen
Kit CB Roes

Contemporary Clinical Trials Communications, 2017; 7, 208-216.

Abstract

Failure to meet subject recruitment targets in clinical trials continues to be a widespread problem with potentially serious scientific, logistical, financial and ethical consequences. On the operational level, enrollment-related issues may be mitigated by careful site selection and by allocating monitoring or training resources proportionally to the anticipated risk of poor enrollment. Such procedures require estimates of the expected recruitment performance that are sufficiently reliable to allow centers to be sensibly categorized. In this study, we investigate whether information obtained from feasibility questionnaires can potentially be used to predict which centers will and which centers will not meet their enrollment targets by means of multivariable logistic regression analysis. From a large set of 59 candidate predictors, we determined the subset that is optimal for predictive purposes using Least Absolute Shrinkage and Selection Operator (LASSO) regularization. Although the extent to which the results are generalizable remains to be determined, they indicate that the prediction accuracy of the optimal model is only a marginal improvement over the intercept-only model, illustrating the difficulty of prediction in this setting.

1. Introduction

Successful completion of a clinical trial requires pre-specified recruitment targets to be met. However, failure to recruit sufficient numbers of subjects in clinical trials continues to be a widespread problem with potentially serious scientific, logistical, financial and ethical consequences [1-5].

On the operational level, enrollment-related issues may be mitigated by careful site selection (i.e., by primarily initiating centers likely to meet enrollment targets and timelines), and by allocating monitoring or training resources proportionally to the anticipated risk of poor enrollment. Often, considerable effort is made to collect information on topics related to enrollment performance by means of extensive feasibility questionnaires. Yet, how to reliably draw an a priori distinction between centers that will and centers that will not meet their enrollment targets, and whether doing so is sensible, is unclear. It requires knowledge concerning potential associations between quantifiable center-specific factors and recruitment performance.

In this study, we investigate such associations. In a broad sense, it shares this aim with many earlier studies (e.g. [4, 6-13]). Note, however, that there exist considerable heterogeneity between these studies in terms of e.g. medical context, methodology, the operationalization of 'recruitment performance', and the results, making it far from clear which factors should be used for the purpose of an operational risk classification.

Prior to the recruitment phase, investigators are typically required to present an enrollment plan, based on their expectations regarding the available number of eligible patients, their willingness to cooperate, available staff members, etc. In this study, we consider enrollment to be successful if this center-specific pre-specified enrollment target is met. To our knowledge, only studies described by Reuter and Esche [9] and Getz [11] use this definition as well: Reuter and Esche [9] assessed the association between meeting enrollment targets and various feasibility questionnaire responses (aimed to measure, among others, the size of the potential patient pool, site/staff experience, and concerns with respect to the investigational medicinal product) in a phase III rheumatoid arthritis clinical trial, but did not

detect any significant associations. Getz summarizes the results of a study that concludes that *"once a particular site has conducted six to 10 clinical trials, that site has a higher likelihood of meeting enrollment targets within the requisite time frame."* [11, p.1].

We use data obtained from a large, international, placebo-controlled, phase III cardiovascular clinical trial to further evaluate possible associations between center-specific factors and recruitment performance. We consider a large set of candidate predictors obtained from the trial's feasibility study. To assess whether there exists a subset of candidate predictors that can help to identify which centers will and which will not meet their enrollment target, we use logistic regression analysis in combination with variable selection through Least Absolute Shrinkage and Selection Operator (LASSO) regularization.

2. The AleCardio trial

The AleCardio trial (ClinicalTrials.gov identifier: NCT01042769) aimed to assess the effect of treatment with Aleglitazar on cardiovascular mortality and morbidity in patients with known or newly diagnosed type 2 diabetes (T2D) who experienced a recent acute coronary syndrome (ACS) event. Patient enrollment in the trial took place between February 2010 and May 2012. In July 2013, the trial was halted due to futility for efficacy and increased rates of safety endpoints. In total, over 7000 patients were included by over 700 sites in Asia Pacific, China, Eastern Europe and Russia, India, Latin and South America, North America and Western Europe. For more detailed accounts of the design and results of this trial, see Lincoff et al. [14, 15].

We use data from 811 centers for which an enrollment target was set and which were actually initiated. Note that not all of these centers ended up enrolling subjects. In fact, 88 centers were closed before the anticipated end of the recruitment period, typically because they failed to enroll any subject.

3. Methods

3.1. Outcome and candidate predictors

The outcome of interest is quantified as the dichotomous variable indicating whether a center met its enrollment target timely (0=no, 1=yes). We treat the 88 centers that were closed early as not having met their target, as we consider the theoretical possibility that these centers (had they not been closed early) would have met their enrollment targets infeasible.

Candidate predictors were obtained from the feasibility questionnaire data, an extensive source of information during the center initiation phase. Information was extracted for all items which may possibly associated with recruitment performance, yielding a total of 56 candidate predictor variables. In addition, three candidate predictors were extracted from the recruitment planning that the centers were required to provide: the expected (i.e., target) number of subjects recruited, the expected number of months required to meet that target, and the anticipated screen failure rate. The 59 candidate predictors can be categorized into seven categories: (1) general center characteristics, (2) staff availability, (3) clinical trial experience, (4) patient pool characteristics, (5) potential/perceived enrollment challenges, (6) recruitment plan and strategies, and (7) contract execution and protocol approval. More details are provided in Appendix A.

3.2. Descriptive analyses

For descriptive purposes, we describe the distribution of the target and actual number of enrolled subjects and calculate the proportion of centers meeting the enrollment target. In addition, we regress the outcome variable (i.e., the dichotomous variable indicating whether a center met its enrollment target) on the full set of candidate predictors using multivariable logistic regression analysis, fitted using quasi-likelihood estimation to account for possible overdispersion.

3.3. Variable selection procedure

We use LASSO regularization [16] to determine the subset of candidate predictors that is optimal in terms of prediction accuracy when regressing the outcome on the candidate predictors through multivariable logistic regression analysis. In the estimation of the

regression coefficients, LASSO regularization enables regression coefficient estimates to be shrunk to exactly zero, thereby realizing variable selection. The amount of shrinkage applied to the regression coefficient estimates, and hence the number of regression coefficient estimates equal to zero, is determined by the value of the tuning parameter λ, with larger values of λ representing more shrinkage. To determine the appropriate value of λ, we first estimate the model's out-of-sample prediction error (in terms of the Brier score) by cross validating a grid of 500 possible λ values through 10-fold cross validation (CV). Two common options for selecting λ are (1) to select the value of λ for which the CV error is minimized, and (2) to use the 1-standard error (1-SE) rule, i.e. to select the largest value of λ for which the CV error is within one SE from the minimum CV error. We apply both strategies. To ensure that all levels of categorical predictors are either in- or excluded from the model, a so-called group LASSO is used, as implemented in the R [17] package 'grplasso' [18]. To account for model-selection instability caused by the random selection of the 10 CV folds, we repeat the CV 20 times and calculate the 95^{th} percentile of the 20 selected λ values (see [19]). For each model, we assess the (range of) CV error values at the selected value of λ. In addition, we determine the (range of) cross validated area under the curve (CV-AUC) values for each model at the selected value of λ to investigate the discriminatory power of the models.

3.4. Missing data handling

Table A1 shows the proportion of missing values for each of the candidate predictors. No values were missing for the outcome variable. Using the R [17] package 'mice' [20], we impute missing values ten times by means of predictive mean matching (for numeric data), logistic regression imputation (for binary data), polytomous regression imputation (for unordered categorical data) or proportional odds regression (for ordinal data), the default options in the mice function. As a consequence, the variable selection procedure is repeated ten times. Note that a complete case analysis was performed to assess the impact of removing centers with missing values (results are described in Appendix C).

4. Results

4.1. Descriptive analyses

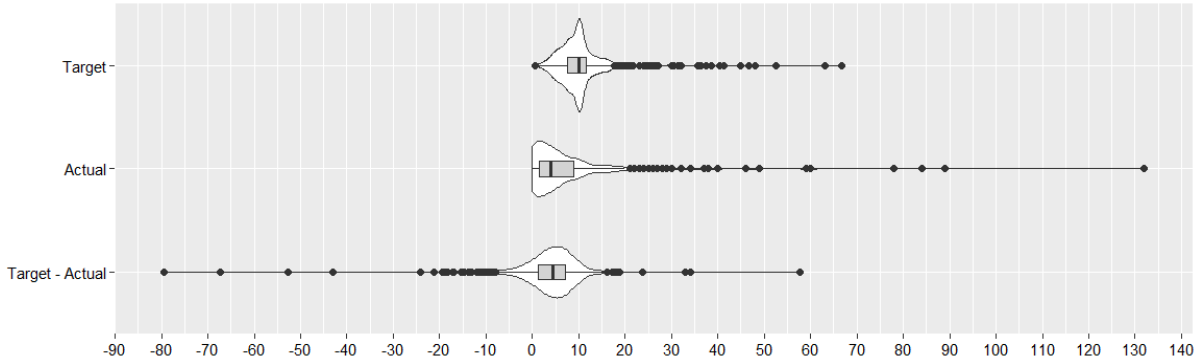

Figure 1. Violin plots and boxplots showing the distribution of (1) the center-specific enrollment targets, (2) the actual number of subjects enrolled, and (3) the difference between the enrollment target and the actual number of subjects enrolled.

Figure 1 shows the distribution of (1) the center-specific enrollment targets, (2) the actual number of subjects enrolled, and (3) the difference between the enrollment target and the actual number of subjects enrolled. The median center-specific enrollment target equals 10.1 (Q1: 7.5, Q3: 11.5). The median number of actually recruited subjects equals 4.0 (Q1: 1.5, Q3: 9.00). It can be seen that only few centers (18.2 percent, 95% Wilson's CI: 15.7 – 21.1) met their enrollment target.

Regressing the outcome (i.e., the variable indicating whether a center met its recruitment target) on the full set of candidate predictors by means of a quasi-binomial generalized linear model and pooling over the multiply imputed datasets yields the results shown in Table 1 (in which, for presentation purposes, only predictors with associated p-values lower than 0.1 are displayed).

Predictor	Description	$\hat{\beta}$	SE	P-value
(Intercept)	-	-4.214	2.198	-
GCC.region	*The region in which a center is located.*			<0.001
Asia Pacific		*(Ref)*	-	
China		1.130	0.626	
E.Europe/Russia		0.021	0.504	
India		1.001	0.674	
Latin/S. America		0.412	0.550	
N. America/Can.		-1.142	0.513	
W. Europe		-0.817	0.489	
GCC.clinic	Indicates whether the center can be considered a clinical setting (0=no, 1=yes)	0.931	0.357	0.003
CTE.distrials_dep	The department's experience (in number of trials) with clinical trials conducted in this disease area.			0.052
None		*(Ref)*	-	
1 to 5		0.657	0.476	
6 to 9		1.248	0.536	
10 or more		1.200	0.538	
PEC.stmed	Indicates whether the center expects the study medication to be a challenge with respect to enrollment. Rated from 1 to 5 with number 1 being the most challenging.	-0.189	0.115	0.072
RPS.alterncontact	Indicates whether the center has stated to be both willing and capable of utilizing alternate contact information for patients, including that of family and friends, to assist in maintaining patient contact (0=no, 1=yes).	-0.548	0.301	0.060
RPS.recr_dur	The planned length (in months) of the follow-up period	0.100	0.032	0.002
RPS.webcasts	Indicates whether the center has stated to be both willing and capable of providing periodic webcasts for patients (0=no, 1=yes).	-1.289	0.628	0.034
CEPA.comm_approv	The total number of days required from submission of essential study documents to obtain final protocol approval from all of the site's required committees combined.			0.036
1 to 10		*(Ref)*	-	
11 to 20		-1.001	0.472	
21 to 30		-0.499	0.409	
31 to 60		-0.706	0.417	
Greater than 60		0.165	0.494	

Table 1. Pooled regression coefficient estimates ($\hat{\beta}$), standard errors (SE), and p-values for the quasi-binomial generalized linear model regressing the outcome (i.e. the variable indicating whether a center met its recruitment target) on the full set of candidate predictors. For presentation purposes, the table only includes predictors associated with p-values lower than 0.1. P-values are based on likelihood ratio tests comparing the full model against the model without the predictor. The estimate of the dispersion parameter ranged from 1.07 to 1.14. See Appendix A for a more detailed description of the variables.

4.2. Predictor selection

Table 1 already provides an indication of which candidate predictors are potentially important, but the results of the more formal variable selection procedure (for each multiply imputed dataset) are presented in Table 2. For instance, when using the 'minimum CV error' strategy to select the shrinkage parameter λ on the first multiply imputed dataset, it can be observed that that the selected value of λ (scaled to the maximum possible value) equals 0.239. At that value of lambda, the CV error ranges[5] from 0.142 to 0.145, the CV-AUC ranges from 0.643 to 0.675, and the selected model contains 13 predictors plus an intercept term. The corresponding CV-plot is provided in Appendix B. It can be observed that, while the CV error and CV-AUC values are similar over the multiply imputed datasets, the set of selected predictors is not, although a subset of eight candidate predictors is selected consistently.

These results, however, do not hold when the 1-SE rule is used to select λ. In that case, the selected value of λ consistently equals its maximum possible value, thus yielding intercept-only models. However, in terms of prediction accuracy, the negative consequences of doing so are marginal, as observed by the limited increase in the CV error and the limited decrease in the CV-AUC.

[5]The CV error and the CV-AUC are variable because the CV procedure was repeated 20 times, as explained in section 3.3.

Strategy for selecting λ	MI dataset	λ (scaled)	CV error	CV-AUC	Selected candidate predictors
Minimum CV error	1	0.239	0.142-0.145	0.643-0.675	**(Intercept), GCC.region**, **GCC.clinic**, **CTE.distrials_dep**, **PPC.num12m**, **PEC.scrfail**, **RPS.recr_target**, **RPS.recr_dur**, **RPS.webcasts**, GCC.inet, CTE.gcptrials_dep, PEC.proc, CEPA.comm_approv, CEPA.exec_30d.
	2	0.229	0.143-0.145	0.632-0.668	**(Intercept), GCC.region**, **GCC.clinic**, **CTE.distrials_dep**, **PPC.num12m**, **PEC.scrfail**, **RPS.recr_target**, **RPS.recr_dur**, **RPS.webcasts**, GCC.medhsp, CTE.gcptrials_dep, PEC.stmed, CEPA.comm_approv, CEPA.exec_30d.
	3	0.313	0.145-0.148	0.611-0.647	**(Intercept), GCC.region**, **GCC.clinic**, **CTE.distrials_dep**, **PPC.num12m**, **PEC.scrfail**, **RPS.recr_target**, **RPS.recr_dur**, **RPS.webcasts**, GCC.smo.
	4	0.235	0.142-0.145	0.645-0.669	**(Intercept), GCC.region**, **GCC.clinic**, **CTE.distrials_dep**, **PPC.num12m**, **PEC.scrfail**, **RPS.recr_target**, **RPS.recr_dur**, **RPS.webcasts**, CTE.gcptrials_dep, PEC.proc, PEC.stmed, PEC.patpop, RPS.chartrev, RPS.promote, RPS.alterncontact, CEPA.comm_approv, CEPA.exec_30d.
	5	0.235	0.144-0.146	0.626-0.675	**(Intercept), GCC.region**, **GCC.clinic**, **CTE.distrials_dep**, **PPC.num12m**, **PEC.scrfail**, **RPS.recr_target**, **RPS.recr_dur**, **RPS.webcasts**, GCC.pi_inv, GCC.medhsp, PEC.stmed, CEPA.comm_approv.
	6	0.236	0.144-0.146	0.630-0.662	**(Intercept), GCC.region**, **GCC.clinic**, **CTE.distrials_dep**, **PPC.num12m**, **PEC.scrfail**, **RPS.recr_target**, **RPS.recr_dur**, **RPS.webcasts**, GCC.inet, GCC.pi_spec, GCC.medhsp, SA.resnurse, CTE.audit, PEC.proc, PEC.stmed, CEPA.comm_approv, CEPA.exec_30d.
	7	0.239	0.144-0.146	0.624-0.650	**(Intercept), GCC.region**, **GCC.clinic**, **CTE.distrials_dep**, **PPC.num12m**, **PEC.scrfail**, **RPS.recr_target**, **RPS.recr_dur**, **RPS.webcasts**, CTE.audit, CTE.gcptrials_dep, CTE.gcpyrs_stcoord, PEC.stmed, CEPA.comm_approv, CEPA.exec_30d.
	8	0.247	0.143-0.146	0.625-0.664	**(Intercept), GCC.region**, **GCC.clinic**, **CTE.distrials_dep**, **PPC.num12m**, **PEC.scrfail**, **RPS.recr_target**, **RPS.recr_dur**, **RPS.webcasts**, GCC.medhsp, SA.recrspec, CTE.audit, CTE.gcpyrs_stcoord, PEC.stmed, PEC.patpop, CEPA.comm_approv, CEPA.exec_30d.
	9	0.231	0.143-0.146	0.626-0.656	**(Intercept), GCC.region**, **GCC.clinic**, **CTE.distrials_dep**, **PPC.num12m**, **PEC.scrfail**, **RPS.recr_target**, **RPS.recr_dur**, **RPS.webcasts**, GCC.inet, GCC.medhsp, GCC.smo, CTE.audit, CTE.gcptrials_dep, PEC.stmed, RPS.chartrev, RPS.alterncontact, CEPA.comm_approv.
	10	0.262	0.144-0.147	0.628-0.660	**(Intercept), GCC.region**, **GCC.clinic**, **CTE.distrials_dep**, **PPC.num12m**, **PEC.scrfail**, **RPS.recr_target**, **RPS.recr_dur**, **RPS.webcasts**, CTE.gcptrials_dep, PEC.stmed, PEC.patpop, RPS.chartrev, CEPA.comm_approv.
1-SE rule	1	1	0.149-0.150	0.484-0.507*	**(Intercept)**
	2	1	0.149-0.150	0.483-0.505*	**(Intercept)**
	3	1	0.149-0.150	0.484-0.504*	**(Intercept)**
	4	1	0.149-0.150	0.490-0.506*	**(Intercept)**
	5	1	0.149-0.150	0.484-0.504*	**(Intercept)**
	6	1	0.149-0.150	0.476-0.503*	**(Intercept)**
	7	1	0.149-0.150	0.486-0.502*	**(Intercept)**
	8	1	0.149-0.150	0.488-0.502*	**(Intercept)**
	9	1	0.149-0.150	0.484-0.500*	**(Intercept)**
	10	1	0.149-0.150	0.492-0.501*	**(Intercept)**

*Table 2. Results (λ, CV error, CV-AUC and the set of selected predictors) of the LASSO analyses for each multiply imputed dataset, using two strategies to select λ. In the last column, variables that are consistently selected are highlighted in bold font. See Appendix A for a description of the variables. *Note that some variability in the results is possible because the maximum value of λ in the CV training sets may not be identical to the maximum value in the complete data.*

The size and direction of the regression coefficient estimates of the candidate predictors that were consistently selected when using the 'minimum CV error' rule for choosing λ are provided in Table 3.

In this study, centers in China and India are predicted to have highest probability of meeting recruitment targets (keeping the other variables constant). Predicted probabilities are lowest for centers in Western Europe and North America and Canada (GCC.region). Higher probabilities are predicted for centers that can be considered clinical settings (GCC.clinic), centers that have more experience with clinical trials in the disease area of interest (CTE.distrials_dep), and for centers with a larger patient pool (PPC.num12m). Also, a higher anticipated screen failure rate positively affects the predicted probability of meeting enrollment targets (PEC.scrfail), as does a longer duration of the recruitment period (RPS.recr_dur). Contrary to our expectations, a positive regression coefficient estimate is found of the enrollment target (RPS.recr_target). A second surprising finding is that centers stating to be both willing and capable of providing periodic webcasts for patients (a strategy aimed to increase patient retention) have lower predicted probabilities to meet their target than center who did not (RPS.webcasts). From the categories 'staff availability (SA)' and 'Contract execution and protocol approval (CETA)', no candidate predictors were consistently selected.

Predictor	**Multiply imputed dataset**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
(Intercept)	-2.648	-2.388	-2.421	-2.409	-2.536	-2.721	-2.437	-2.456	-2.669	-2.499
GCC.region										
Asia Pacific	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*
China	0.614	0.602	0.391	0.593	0.611	0.601	0.570	0.549	0.612	0.504
E.Europe/Russia	0.119	0.105	0.080	0.119	0.096	0.125	0.100	0.105	0.101	0.098
India	0.454	0.432	0.241	0.423	0.451	0.439	0.400	0.392	0.457	0.345
Latin/S. America	0.191	0.156	0.093	0.161	0.149	0.172	0.139	0.144	0.159	0.124
N. America/Can.	-0.290	-0.332	-0.236	-0.303	-0.354	-0.314	-0.321	-0.314	-0.356	-0.289
W. Europe	-0.140	-0.147	-0.104	-0.139	-0.162	-0.144	-0.141	-0.134	-0.175	-0.126
GCC.clinic	0.342	0.425	0.168	0.313	0.366	0.320	0.334	0.372	0.313	0.264
CTE.distrials_dep										
None	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*
1 to 5	0.151	0.155	0.029	0.136	0.181	0.160	0.139	0.148	0.184	0.128
6 to 9	0.359	0.367	0.063	0.339	0.398	0.316	0.307	0.313	0.405	0.246
10 or more	0.294	0.341	0.055	0.316	0.346	0.267	0.258	0.251	0.331	0.245
PPC.num12m	0.137	0.020	0.061	0.025	0.012	0.017	0.002	0.027	0.066	0.039
PEC.scrfail	0.416	0.492	0.394	0.467	0.462	0.469	0.500	0.476	0.489	0.450
RPS.recr_target	0.003	0.004	0.005	0.003	0.004	0.004	0.004	0.004	0.004	0.004
RPS.recr_dur	0.067	0.068	0.060	0.066	0.067	0.066	0.066	0.065	0.068	0.064
RPS.webcasts	-0.617	-0.704	-0.343	-0.873	-0.350	-0.667	-0.415	-0.591	-0.332	-0.481

Table 3. Regression coefficient estimates of the candidate predictors that were consistently selected in each of the multiply imputed datasets. See Appendix A for a description of the variables.

When the ‘1-SE rule’ was used to select λ, the estimated intercept (the only term in the models) was equal to -1.5 in all analyses.

5. Discussion

We used data from a large international phase III cardiovascular clinical trial to investigate associations between center characteristics obtained from the responses to the feasibility questionnaires and meeting recruitment targets. From a large number of candidate predictors, we determined the subset that is optimal for predictive purposes using LASSO regularization. In terms of prediction accuracy, the models selected using the ‘minimum CV error’ strategy for choosing the value of shrinkage parameter λ are only marginal improvements over the intercept-only models that were selected using the ‘1-SE rule’. This result implies that the predictive value of the set of candidate predictors is limited and should not be overestimated. The results illustrate the difficulty of prediction in this setting and suggest that, in this context, it may be unjustified to base operational decisions on the responses to the feasibility questionnaire items.

However, these findings are based on data from a single trial, hampering their generalizability. In addition, using the results of our study specifically for the purpose of making a decision to proceed or not to proceed with a center in the site selection process should be done with caution, as the selection of centers included in this study already represents a (possibly selective) subset of centers. More research is needed to assess whether the findings presented here hold more generally.

The data used were not collected for the purpose of this analysis. Therefore, this assessment should be considered explorative in nature. We were unable to include potentially relevant feasibility questionnaire items due to, e.g., ambiguous item or answer formulations, and in some cases a subjective assessment of text field entries was required. Data on certain potentially important factors were not collected. E.g. our list of candidate predictors fails to adequately address the extent and nature of potential prior cooperation between the center and the sponsor or site management organization. Also, the feasibility questionnaire was designed

specifically for this trial and, as a consequence, questionnaire items were not always formulated in sufficiently general terms. Lastly, one could argue that a formal comparison of the predictive accuracy of the model constructed using the 'minimum CV error' rule versus the model based on the '1-SE rule' requires independent test data. We therefore repeated the LASSO procedure on a random selection of two-thirds of the data, then estimated the prediction error and AUC values on the remainder of the data. Although the results (available upon request) showed signs of numerical instability (i.e. selected λ values were more variable over the multiply-imputed data sets, likely due to the smaller sample sizes of the training sets), the prediction error levels and AUC values corresponding to the selected models are similar to the results described above.

Comparing our results to the results of earlier investigations is not straightforward due to variability in terms of medical context, study methodology and the operationalizations used. Note, however, that from a general perspective our results resemble those of Reuter and Esche [9] who failed to detect a significant association between the responses to a range of feasibility questionnaire items and meeting enrollment targets. These results reveal possible limitations of the items used in feasibility questionnaires and could be interpreted as a warning against overemphasizing the outcomes of feasibility studies in general. Overall, however, more research is needed to be able to draw more definitive conclusions. The candidate predictors selected using the 'minimum CV error' strategy for choosing λ may have been of limited value in this trial, but may be considered for re-evaluation in future studies.

In conclusion, the results suggest that drawing a reliable a priori distinction between centers that will meet their recruitment target and those that will not is a difficult task, as even the optimal selection of candidate predictors only represents a marginal improvement in predictive accuracy as compared to the intercept-only model. Thus, the predictive value of current feasibility studies may not be large enough to justify such extensive questioning. However, more research, preferably from varying types of trials and clinical contexts, is needed to assess whether our results hold more general, or whether there may be other factors that need to be included.

Appendix A: List of candidate predictors

Candidate predictor	Description	% missing
General center characteristics (GCC)		
GCC.emr	Indicates whether the center has access to electronic medical records (0=no, 1=yes). Percentages: 41.2, 58.8.	3.9
GCC.inet	Indicates whether a high-speed Internet connection is available at the center (0=no, 1=yes). Percentages: 1.9, 98.1.	3.9
GCC.pdb	Indicates whether the center has access to a patient database (0=no, 1=yes). Percentages: 20.9, 79.1.	3.9
GCC.fu_resp	Indicates whether the center is responsible for the long-term follow-up of patients in this trial (0=no, 1=yes). Percentages: 6.7, 93.3.	6.2
GCC.pi_inv	Indicates whether the PI is routinely involved in follow-up visits with study patients (0=no, 1=yes). Percentages: 9.2, 90.8.	6.2
GCC.pi_spec	Indicates whether the PI's specialty corresponds to the research area (here, cardiology/diabetes). (0=no, 1=yes). Percentages: 2.9, 97.1.	2.1
GCC.region	The region in which a center is located. Similar to Desai et al. [21], each site is classified into one of the following regions: Asia Pacific, China, Eastern Europe and Russia, India, Latin and South America, North America (United States and Canada), Western Europe. Percentages: 9.2, 4.4, 12.5, 4.7, 13.1, 35.4, 20.7.	0
GCC.clinic	Indicates whether the center can be considered a clinical setting (0=no, 1=yes). Percentages: 88.5, 11.5.	0.5
GCC.crc	Indicates whether the center can be considered a clinical research center (0=no, 1=yes). Percentages: 85.6, 14.4.	0.5
GCC.gov	Indicates whether the center can be considered a government-run medical facility (0=no, 1=yes). Percentages: 89.3, 10.7.	0.5
GCC.group	Indicates whether the center can be considered a group practice (0=no, 1=yes). Percentages: 85.5, 14.5.	0.5
GCC.medhsp	Indicates whether the center can be considered a medical hospital (0=no, 1=yes). Percentages: 51.7, 48.3.	0.5
GCC.private	Indicates whether the center can be considered a private practice (0=no, 1=yes). Percentages: 75.2, 24.8.	0.5
GCC.smo	Indicates whether the center can be considered a site management organization (0=no, 1=yes). Percentages: 97.6, 2.4.	0.5
GCC.spec	Indicates whether the center can be considered a cardiology specialist center (0=no, 1=yes). Percentages: 98.6, 1.4.	0.5
GCC.teach	Indicates whether the center can be considered a teaching hospital (0=no, 1=yes). Percentages: 85.4, 14.6.	0.5
Staff availability (SA)		
SA.diet	Indicates whether a registered dietician/nutritionist is available (1=yes, 0=no). Percentages: 62.2, 37.8.	6.4
SA.endocr	Indicates whether an endocrinologist is available (1=yes, 0=no). Percentages: 59.6, 40.4.	6.4
SA.pharm	Indicates whether a pharmacologist is available (1=yes, 0=no). Percentages: 51.6, 48.4.	6.4
SA.phleb	Indicates whether a phlebotomist is available (1=yes, 0=no). Percentages: 59.6, 40.4.	6.4
SA.radiol	Indicates whether a radiologist is available (1=yes, 0=no). Percentages: 58.5, 41.5.	6.4
SA.recrspec	Indicates whether a recruitment specialist is available (0=no, 1=yes). Percentages: 81.0, 19.0.	6.4
SA.resnurse	Indicates whether a research nurse is available (0=no, 1=yes). Percentages: 30.7, 69.3.	6.4
SA.stcoord	Indicates whether a study coordinator is available (0=no, 1=yes). Note: if missing or 0, but CTE.gcpyrs_stcoord is > 0, set to 1. Percentages: 5.7, 94.3.	6.4
SA.subi	Indicates whether a sub-investigator is available (0=no, 1=yes). Note: if missing or 0, but CTE.gcpyrs_subi is > 0, set to 1. Percentages: 6.9, 93.1.	6.4
Clinical trial experience (CTE)		
CTE.audit	Indicates whether the center has ever been audited by a regulatory agency or health authority (0=no, 1=yes). Percentages: 80.7, 19.3.	6.8
CTE.gcptrials_dep	The department's experience (number of trials in the past three years) with clinical trials conducted according to ICH and GCP Guidelines. Categories: None, 1 to 4, 5 to 9, and 10 or more. Percentages: 1.3, 14.5, 30.2, 54.1.	4.4
CTE.distrials_dep	The department's experience (in number of trials) with clinical trials conducted in this disease area. Categories: None, 1 to 5, 6 to 9, and 10 or more. Percentages: 11.3, 45.2, 19.2, 24.2.	4.3
CTE.gcpyrs_pi	The PI's experience (in years) with clinical trials conducted according to ICH and GCP Guidelines. Categories: None, less than 1 year, 1 to 4 years, 4 to 7 years, or greater than 7 years. Percentages: 1.5, 2.1, 11.0, 20.5, 64.9.	4.3

CTE.gcpyrs_stcoord	The study coordinator's experience (in years) with clinical trials conducted according to ICH and GCP Guidelines. Categories: None, less than 1 year, 1 to 4 years, 4 to 7 years, or greater than 7 years. Equals 0 if no study coordinator is present. Percentages: 6.5, 4.5, 22.5, 26.3, 40.2.	7.3
CTE.gcpyrs_subi	The sub-investigator's experience (in years) with clinical trials conducted according to ICH and GCP Guidelines. Categories: None, less than 1 year, 1 to 4 years, 4 to 7 years, or greater than 7 years. Equals 0 if no sub-investigator is present. Percentages: 7.8, 6.0, 23.1, 27.1, 36.0.	7.2

Patient pool characteristics(PPC)

PPC.patdis10km	What proportion of your patients live within approximately 10 km (6 miles) distance from your clinic? 0, .01-.2, .21-.4, .41-.6, .61-.8, or >.8? Used category midpoints, treated as continuous. Q1, Q2, Q3: 0.30, 0.50, 0.70.	8.3
PPC.num12m	The number of ACS patients with newly diagnosed T2D the center treated during the past 12 months, divided by 100. Note that this is an approximation, as it is based on two questions (one for ACS, and one for T2D, with ordinal answer categories). The product of midpoints was used. Furthermore, since one of the two questions had an open-ended last category, the strategy described and recommended by Parker & Fenwick [22] was used to estimate the midpoint for this category. Note also that this item excludes ACS patients with known T2D. Q1, Q2, Q3: 0.23, 0.49, 1.16.	6.9

Potential or perceived enrollment challenges (PEC)

PEC.proc	Do you expect the procedures or assessments required to be a challenge with respect to enrollment? Please rate from 1 to 5 (with number 1 being the most challenging). Treated as continuous. Q1, Q2, Q3: 3, 4, 5.	6.8
PEC.import	Do you expect the importation issues to be a challenge with respect to enrollment? Please rate from 1 to 5 (with number 1 being the most challenging). Treated as continuous. Q1, Q2, Q3: 3, 4, 5.	8.1
PEC.inex	Do you expect in- and exclusion criteria to be a challenge with respect to enrollment? Please rate from 1 to 5 (with number 1 being the most challenging). Q1, Q2, Q3: 3, 4, 4.	6.8
PEC.stmed	Do you expect the study medication to be a challenge with respect to enrollment? Please rate from 1 to 5 (with number 1 being the most challenging). Treated as continuous. Q1, Q2, Q3: 3, 4, 5.	7.2
PEC.reimb	Do you expect medication reimbursement issues to be a challenge with respect to enrollment? Please rate from 1 to 5 (with number 1 being the most challenging). Treated as continuous. Q1, Q2, Q3: 3, 4, 5.	7.8
PEC.patpop	Do you expect the patient population to be a challenge with respect to enrollment? Please rate from 1 to 5 (with number 1 being the most challenging). Treated as continuous. Q1, Q2, Q3: 3, 4, 4.	6.8
PEC.regul	Do you expect regulatory authority issues to be a challenge with respect to enrollment? Please rate from 1 to 5 (with number 1 being the most challenging). Treated as continuous. Q1, Q2, Q3: 3, 4, 5.	7.4
PEC.staff	Do you expect a lack of sufficient staff resources to be a challenge with respect to enrollment? Please rate from 1 to 5 (with number 1 being the most challenging). Treated as continuous. Q1, Q2, Q3: 4, 5, 5.	7.3
PEC.visit_dur	Do you expect the visit frequency and/or study duration to be a challenge with respect to enrollment? Please rate from 1 to 5 (with number 1 being the most challenging). Treated as continuous. Q1, Q2, Q3: 3, 4, 5.	6.9
PEC.impconcerns	Indicates whether the center has concerns about the investigational medicinal product (0=no, 1=yes). Percentages: 79.7, 20.3.	6.0
PEC.comptrials	Indicates whether other, possibly competing trials are currently running/planned on the center. Categories: "No", "Yes, but we can still meet the enrollment goal for this study", "Yes, and it may impact our ability to meet the enrollment goal for this study". Percentages: 60.1, 35.9, 4.0.	7.3
PEC.scrfail	The expected proportion of screen failures. Q1, Q2, Q3: 0.15, 0.20, 0.25.	0

Recruitment plan and strategies (RPS)

RPS.recr_target	The planned/target number of enrolled subjects. Q1, Q2, Q3: 7.50, 10.05, 11.51.	0
RPS.recr_dur	The planned length (in months) of the follow-up period. Q1, Q2, Q3: 9, 12, 15.	0
RPS.chartrev	Indicates whether the center has stated to be both willing and capable of providing additional support to assist with chart review to identify patients for the study (0=no, 1=yes). Percentages: 59.0, 41.0.	8.0
RPS.promote	Indicates whether the center has stated to be both willing and capable of providing materials or services to promote the study to referral physicians/other departments (0=no, 1=yes). Percentages: 52.1, 47.9.	8.0
RPS.contact	Indicates whether the center has stated to be both willing and capable of to keep regular contact between visits (0=no, 1=yes). Percentages: 42.2, 57.8.	7.4
RPS.cfu_remind	Indicates whether the center has stated to be both willing and capable of providing community follow-up and visit reminder emails, cards and phone calls (0=no, 1=yes). Percentages: 54.6, 45.4.	7.4

RPS.contact_caregiver	Indicates whether the center has stated to be both willing and capable of maintaining contact with the patients' other caregivers, particularly primary care physicians (0=no, 1=yes). Percentages: 56.1, 43.9.	7.5
RPS.letter	Indicates whether the center has stated to be both willing and capable of providing personal thank you letters to patients (0=no, 1=yes). Percentages: 63.9, 36.1.	7.4
RPS.alterncontact	Indicates whether the center has stated to be both willing and capable of utilizing alternate contact information for patients, including that of family and friends, to assist in maintaining patient contact (0=no, 1=yes). Percentages: 66.0, 34.0.	7.4
RPS.items	Indicates whether the center has stated to be both willing and capable of providing study-pertinent items to patients at milestone visits (i.e. diabetes recipes, exercise guides, etc.) (0=no, 1=yes). Percentages: 46.1, 53.9.	7.5
RPS.website	Indicates whether the center has stated to be both willing and capable of creating a study community website for patients to view news and articles related to their condition (0=no, 1=yes). Percentages: 78.8, 21.2.	7.4
RPS.webcasts	Indicates whether the center has stated to be both willing and capable of providing periodic webcasts for patients (0=no, 1=yes). Percentages: 90.4, 9.6.	7.4
Contract execution and protocol approval (CEPA)		
CEPA.comm_approv	The total number of days required from submission of essential study documents to obtain final protocol approval from all of the site's required committees combined. Categories: 1 to 10, 11 to 20, 21 to 30, 31 to 60, Greater than 60. Percentages: 15.4, 15.1, 27.8, 30.8, 10.8.	7.9
CEPA.exec_30d	Indicates whether it usually takes the center more than 30 days to execute a contract and budget (0=yes or unknown, 1=no). Percentages: 70.5, 29.5.	7.8

Table A1. Description of the candidate predictors considered in the analysis. Summary statistics (Quantiles Q1, Q2 and Q3 or percentages) are calculated on observed data. Abbreviations: PI = Principal Investigator, ICH = International Conference of Harmonization, GCP = Good Clinical Practice, ACS = Acute Coronary Syndrome, T2D=Type 2 Diabetes.

Appendix B: Example cross-validation plot

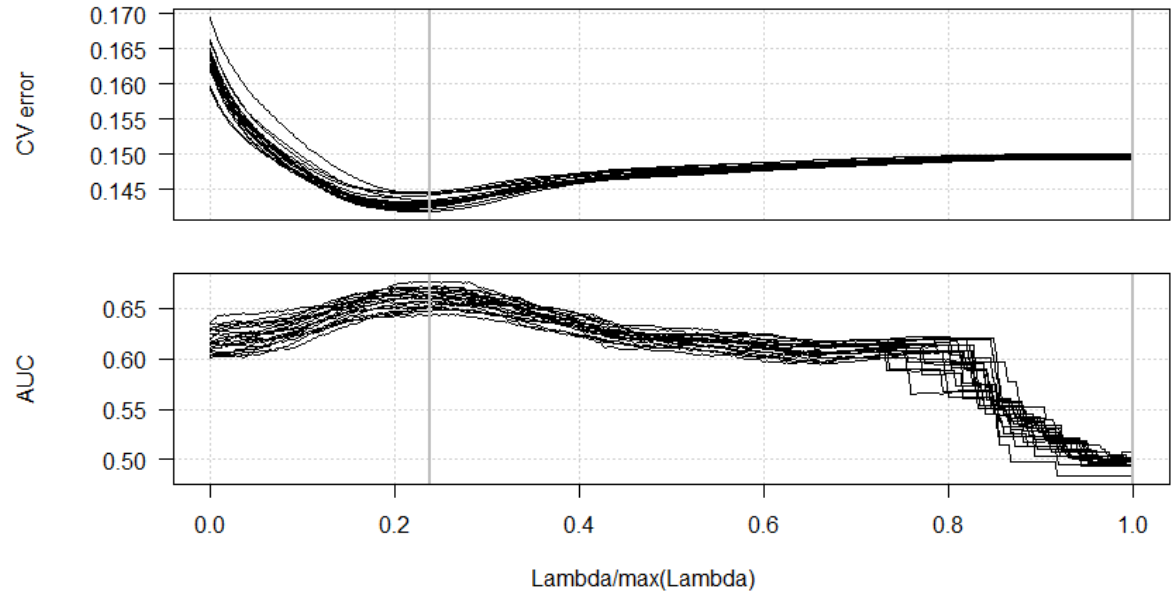

Figure B1. The 20 CV plots for the CV error and the CV-AUC for the analysis performed on the first multiply imputed dataset. The vertical grey lines indicate the selected values of λ when using the minimum CV-error rule (left) or the 1-SE rule (right).

Appendix C: Complete case analysis

For reference, the table below shows the results for the regression analysis described in section 4.1 when it is applied to the 672 centers with complete data. Again, for presentation purposes, only results for which $p < 0.1$ are shown.

Predictor	Description	$\hat{\beta}$	SE	P-value
(Intercept)	-	-4.108	2.404	-
GCC.inet	*Indicates whether a high-speed Internet connection is available at the center (0=no, 1=yes).*	1.781	1.045	0.062
GCC.region	*The region in which a center is located.*			<0.001
Asia Pacific		*(Ref)*	-	
China		1.635	0.676	
E.Europe/Russia		-0.142	0.520	
India		0.924	0.699	
Latin/S. America		0.039	0.587	
N. America/Can.		-1.535	0.569	
W. Europe		-0.823	0.515	
GCC.clinic	*Indicates whether the center can be considered a clinical setting (0=no, 1=yes)*	0.738	0.410	0.077
CTE.gcptrials_dep	*The department's experience (number of trials in the past three years) with clinical trials conducted according to ICH and GCP Guidelines.*			0.032
None		*(Ref)*	-	
1 to 4		-3.403	1.895	
5 to 9		-3.749	1.942	
10 or more		-4.290	1.944	
CTE.distrials_dep	*The department's experience (in number of trials) with clinical trials conducted in this disease area.*			0.041
None		*(Ref)*	-	
1 to 5		0.591	0.510	
6 to 9		1.288	0.564	
10 or more		1.270	0.577	
PEC.proc	*Indicates whether the center expects the procedures or assessments required to be a challenge with respect to enrollment. Rated from 1 to 5 with number 1 being the most challenging.*	-0.279	0.165	0.090
RPS.recr_dur	*The planned length (in months) of the follow-up period*	0.096	0.035	0.006
RPS.webcasts	*Indicates whether the center has stated to be both willing and capable of providing periodic webcasts for patients (0=no, 1=yes).*	-1.721	0.724	0.006
CEPA.comm_approv	*The total number of days required from submission of essential study documents to obtain final protocol approval from all of the site's required committees combined.*			0.010
1 to 10		*(Ref)*	-	
11 to 20		-1.452	0.504	
21 to 30		-0.735	0.441	
31 to 60		-1.012	0.446	
Greater than 60		-0.166	0.524	

Table C1. Complete case analysis: Regression coefficient estimates ($\hat{\beta}$), standard error (SE), and p-values for the quasi-binomial generalized linear model regressing the outcome (i.e. the variable indicating whether a center met its recruitment target) on the full set of candidate predictors. For presentation purposes, the table only includes predictors associated with p-values lower than 0.1. P-values are based on likelihood ratio tests comparing the full model against the model without the predictor. The estimate of the dispersion parameter equals 1.081. See Appendix A for a more detailed description of the variables.

The LASSO procedure, when applied to the subset of centers with complete data, yields the following results: The λ value that minimizes the CV error, as a proportion of its maximum possible value, equals 0.227, at which the CV error ranges from 0.146 to 0.150 and the CV-AUC ranges from 0.629 to 0.680. Using the '1-SE rule' yields a CV error estimate of 0.154 in all settings and a CV-AUC ranging from 0.479 to 0.505. These results are comparable to the results of the analyses on the multiply imputed data. Again, using the '1-SE rule' implies that the optimal model is the intercept-only model (with an intercept of -1.457). Minimizing the CV error results in the model in Table C2. As can be seen, the model contains the set of candidate predictors that was also consistently selected in the analyses on the imputed data, with comparable regression coefficient estimates.

Predictor	**Estimate**
(Intercept)	-2.351
GCC.region	
Asia Pacific	(Ref)
China	0.763
E.Europe/Russia	0.165
India	0.395
Latin/S. America	0.101
N. America/Can.	-0.357
W. Europe	-0.088
GCC.clinic	0.163
GCC.group	-0.100
SA.stcoord	0.128
CTE.gcptrials	
None	(Ref)
1 to 4	-0.237
5 to 9	-0.195
10 or more	-0.323
CTE.distrials	
None	(Ref)
1 to 5	0.145
6 to 9	0.394
10 or more	0.364
PPC.num12m	0.056
PEC.proc	-0.033
PEC.patpop	0.002
PEC.scrfail	0.432
RPS.recr_target	0.006
RPS.recr_dur	0.068
RPS.chartrev1	0.002
RPS.webcasts1	-0.658
CEPA.comm_approv	
1 to 10	(Ref)
11 to 20	-0.212
21 to 30	-0.137
31 to 60	-0.152
Greater than 60	0.164

Table C2. Results (selected candidate predictors and regression coefficient esitmates) of the LASSO procedure when applied to the subset of complete cases and when selecting the value for λ which minimizes the CV error. See Appendix A for a more detailed description of the variables.

References

1. Lovato LC, Hill K, Hertert S, Hunninghake DB, Probstfield JL. Recruitment for controlled clinical trials: Literature summary and annotated bibliography. Control Clin Trials 1997; l8:328-357.
2. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, Elbourne DR, Francis D, Garcia J, Roberts I, Snowdon C. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials 2006; 7:9. DOI: 10.1186/1745-6215-7-9.
3. Friedman LM, Furberg CD, DeMets, DL. Fundamentals of Clinical Trials. Fourth Edition. Springer 2010. DOI: 10.1007/978-1-4419-1586-3.
4. Rahman S, Majumder AA, Shaban SF, Rahman N, Ahmed M, Abdulrahman KB, D'Souza, UJA. Physician participation in clinical research and trials: issues and approaches. Adv Med Educ and Pract 2011;2:85–93. DOI: 10.2147/AMEP.S14103.
5. Fletcher B, Gheorghe A, Moore D, Wilson S, Damery S. Improving the recruitment activity of clinicians in randomized controlled trials: a systematic review. BMJ Open 2012; 2:e000496. DOI: 10.1136/bmjopen-2011-000496.
6. Benson III AB, Pregler JP, Bean JA, Rademaker AW, Eshler B, Anderson K. Oncologists' reluctance to accrue patients onto clinical trials: An Illinois cancer center study. J Clin Oncol 1991; 9(11):2067–2075.
7. Shea S, Bigger JT, Campion J, Fleiss JL, Rolnitzky LM, Schron E, Gorkin L, Handshaw K, Kinney MR, Branyon M. Enrollment in clinical trials: Institutional factors affecting enrollment in the Cardiac Arrhythmia Suppression Trial (CAST). Control Clin Trials 1992: 13;466–486.
8. Donovan JL, Peters TJ, Noble S, Powell P, Gillatt D, Oliver SE, Lane JA, Neal DE, Hamdy FC, for the ProtecT Study Group. Who can best recruit to randomized trials? Randomized trial comparing surgeons and nurses recruiting patients to a trial of treatments for localized prostate cancer (the ProtecT study). J Clin Epidemiol 2003; 56:605–609. DOI: 10.1016/S0895-4356(03)00083-0.
9. Reuter S, Esche G. How effective are site questionnaires in predicting site performance? J Clin Res Best Pract 2007; 3(4).

10. Ford E, Jenkins V, Fallowfield L, Stuart N, Farewell D, Farewell V. Clinicians' attitudes towards clinical trials of cancer therapy. Br J Cancer 2011; 104(10):1535–1543. DOI: 10.1038/bjc.2011.119. Epub 2011 Apr 12.
11. Getz K. Predicting Successful Site Performance. Applied Clinical Trials, November 1st, 2011.
12. Levett M, Roberts CL, Simpson JM, Morris JM. Site-specific predictors of successful recruitment to a perinatal clinical trial. Clin Trials 2014; 11(5):584–589. DOI: http://dx.doi.org/10.1177/1740774514543539.
13. Oude Rengerink K, Hooft L, Van den Boogaard NM, Van der Goes BY, Bossuyt PMM, Mol BWJ. Why do some centres recruit better than others? An analysis of recruitment rates in 17 randomized clinical trials in obstetrics and gynaecology. Chapter 4 in 'Embedding trials in evidence-based clinical practice' by Oude Rengerink, K, PhD thesis, University of Amsterdam, the Netherlands, 2014.
14. Lincoff AM, Tardif J-C, Neal B, Nicholls SJ, Rydén L, Schwartz GG, Malmberg K, Buse JB, Henry RR, Wedel H, Weichert A, Cannata R, Grobbee DE. Evaluation of the dual peroxisome proliferator-activated receptor α/γ agonist aleglitazar to reduce cardiovascular events in patients with acute coronary syndrome and type 2 diabetes mellitus: rationale and design of the AleCardio trial. Am Heart J 2013; 166(3):429–434. DOI: 10.1016/j.ahj.2013.05.013.
15. Lincoff AM, Tardif J-C, Schwartz GG, Nicholls SJ, Rydén L, Neal B, Malmberg K, Wedel H, Buse JB, Henry RR, Weichert A, Cannata R, Svensson A, Volz D, Grobbee DE. Effect of aleglitazar on cardiovascular outcomes after acute coronary syndrome in patients with type 2 diabetes mellitus: The AleCardio randomized clinical trial. JAMA 2014; 311(15):1515–1525. DOI: 10.1001/jama.2014.3321.
16. Tibshirani R. Regression shrinkage and selection via the LASSO. J Royal Stat Soc (Series B), 1996; 58(1):267–288.
17. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016. URL: http://www.r-project.org/. Version 3.3.1.
18. Meier L. grplasso: Fitting user specified models with Group Lasso penalty. 2015. URL: https://CRAN.R-project.org/package=grplasso. Version 0.4-5.

19. Roberts S, Nowak G. Stabilizing the LASSO against cross-validation variability. Comput Stat Data Anal 2014; 70:198–211. DOI: 10.1016/j.csda.2013.09.008.
20. Van Buuren, S, Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. J Stat Soft 2011; 45(3):1-67. URL: http://www.jstatsoft.org/v45/i03/. Version 2.30.
21. Desai PB, Anderson C, Sietsema WK. A comparison of the quality of data, assessed using query rates, from clinical trials conducted across developed versus emerging global regions. Drug Inf J 2012; 46:455–463. DOI: 10.1177/0092861512446807.
22. Parker RN, Fenwick R. The Pareto curve and its utility for open-ended income distributions in survey research. Soc Forces 1983; 61(3):872–885. DOI: 10.2307/2578140.

Chapter 5

The impact of outcome misclassification and the value of adjudication in multi-center clinical trials

Rutger M van den Bor
Kit CB Roes
Petrus WJ Vaessen
Bas J Oosterman
Diederick E Grobbee

Submitted

Abstract

Adjudication committees in clinical trials aim to detect and correct potential errors in the diagnoses of clinical events as made by local investigators. However, including an adjudication process in multi-center, double blinded trials will generally require substantial investments in time and resources. Therefore, the decision regarding the implementation of adjudication in a trial requires careful consideration. This paper aims to aid trialists in this respect, by (1) quantifying the impact of misdiagnosis of trial outcome events on estimates of incidences and treatment effects, as well as its effect on the power of statistical tests, and (2) discussing the extent to which adjudication addresses those issues. We show how misdiagnoses has the potential to invalidate trial results, even if the probability of misdiagnosis is comparable in the two trial arms. Adjudication may, in practice, be necessary to ensure the validity of the results.

1. Introduction

In the conduct of clinical trials, 'adjudication' refers to the process by which the diagnoses of important clinical events, as made by local investigators, are reviewed by a central committee (typically blinded for treatment assignment) with the aim of detecting and correcting potential cases of misdiagnosis. This process is commonly used when the definition of the event is heterogeneous or to some extent subjective (e.g. myocardial infarction, stroke or cardiovascular/vascular death in cardiovascular trials) ([1-3]). It is believed that adjudication adds to the reliability of the trial findings by increasing the precision of the results and by reducing potential biases [3]. However, including an adjudication process in a trial will generally require substantial investments in time and resources [2-5]. We therefore agree with Hata et al., who state that the *"appointment of an [adjudication committee] […] warrants the same careful scientific consideration as other aspects of the trial design"* [4, p.7]. This paper aims to aid trialists in this respect, by (1) quantifying the impact of misdiagnosis of trial outcome events on estimates of incidences and treatment effects, as well as its effect on the power of statistical tests, and (2) discussing the extent to which adjudication addresses those issues. We do so in the context of multi-center, double blinded trials.

2. The impact of misdiagnosis of trial outcome events

The act of diagnosing can be considered a classification process, the accuracy of which is defined in terms of its sensitivity, the probability of correctly identifying an event or case, and its specificity, the probability of correctly identifying a non-event or non-case.

The impact of misdiagnosis depends on whether the misclassification is differential (i.e. 'non-random', meaning that structural differences exist in terms of the sensitivity and specificity of the classification process between the two treatment arms) or not. In general, misclassification in clinical trials can be assumed to be non-differential ('random'), in particular when double blinding is used. Although it is known that non-differential misdiagnosis of outcome events generally will attenuate effect estimates [6], a recent study concludes that *"a substantial amount of random errors is required before appreciable effects on the outcome of randomized clinical trials are noted."* [7, p.1]. In addition, the impact of misdiagnosis depends on the type of estimator that is used. E.g. different types of estimators of the treatment effect are

differentially affected by non-differential misclassification. For instance, it is known that the risk ratio remains unbiased as long as the specificity equals one - a property that does not hold for, e.g., the risk difference.

In what follows, we describe and graphically illustrate the impact of non-differential misdiagnosis of a binary outcome on the results of multi-center clinical trials. We start by investigating the impact of misclassification in terms of exposure-specific incidence estimates. Next, we consider the impact on estimates of effects (the risk difference (RD), the risk ratio (RR), the odds ratio (OR) and the hazard ratio (HR)), and subsequently we discuss the impact of misclassification on the statistical power. The assessments are based on a simulation procedure, in which sensitivity and specificity levels are varied in multiple simulated scenarios. The simulation algorithms are detailed in Appendix A.

Note that, in multi-center trials, it can be expected that sensitivity and specificity levels vary from center to center, due to differences in e.g. training, setting, or standard of care. In our simulations, this is taken into account by generating distributions of sensitivity and specificity values instead of fixed constants.

2.1. Estimation of the sample incidence

Suppose an investigation aims to estimate parameter p, the probability of a subject developing the outcome of interest during a particular fixed time period (i.e. the incidence), using a (representative) sample of n subjects. Let y_i indicate the presence (1) or absence (0) of the outcome in subject i, and assume it can be observed without any error or misclassification. If the observations are independent and subject to the same probability of having an event, the sum of y_i over the n subjects is binomially distributed. The incidence p is estimated by $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$, so that $E(\hat{p}) = p$ and $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, the latter typically being estimated by $\widehat{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

This situation, however, is actually theoretical: Only in the absence of any error or misclassification is y_i directly observable. Under non-perfect classification, the apparent

event status (denoted by y_i^*) is observed instead. The apparent sample incidence then equals $\hat{p}^* = \frac{1}{n}\sum_{i=1}^{n} y_i^*$. If the misclassification probabilities (i.e. $1-\theta$ and $1-\tau$) are the same for all n subjects , the sum of y_i^* is still binomially distributed, but $E(\hat{p}^*) = \theta p + (1-\tau)(1-p)$ and therefore $SE_{\hat{p}}^* = \sqrt{\frac{E(\hat{p}^*)(1-E(\hat{p}^*))}{n}}$ [8]. That is, a statistical bias is introduced, the size and direction of which depend on the particular combination of sensitivity θ, specificity τ and the true incidence.

The diagnostic accuracy by which events are classified will, in multi-center investigations, usually not be constant over centers. Therefore, let θ_j and τ_j denote the sensitivity and specificity of the diagnosis of y_i in the jth center. Within each center, $E(\hat{p}_j^*) = \theta_j p + (1-\tau_j)(1-p)$. As a result, the overall estimate of $\hat{p}^*$ will again be biased. Note that the additional variability due to variable sensitivity and specificity also introduces overdispersion, complicating the estimation of the standard error of $\hat{p}^*$.

Figure 1 illustrates these observations in more detail: The first row shows the expected value of $\hat{p}^*$ for p equal to 0.125, 0.5, and 0.875 and for a range of median sensitivity and specificity levels in a simulated multi-center study with 5 centers and 50 subjects per center. Although very low sensitivity and specificity values may be rare in practice, we chose to consider the full range of possible values, because there may be considerable heterogeneity in the diagnostic accuracy for different types of clinical outcomes. It can be observed that, in all scenarios, the expected value $E(\hat{p}^*)$ can extend through the whole range from 0 to 1, hence any level of bias can in principle occur. When the median sensitivity and specificity both equal 1 (i.e. when classification is perfect), the expected value equals the true incidence $E(\hat{p}^*) = p$. When both medians equal 0, $E(\hat{p}^*) = 1-p$. When the median sensitivity and specificity values sum to 1, $E(\hat{p}^*)$ is approximately equal to the median sensitivity. For larger values of p, the relative importance of the sensitivity over the specificity increases.

The second row shows the impact of ignoring the overdispersion due to between-center variability of misclassification rates in the estimation of the standard error. Specifically, the ratio of the mean value of the naively applied standard estimate of the standard error $\widehat{SE}_{\hat{p}}$ and

the empirical standard error (i.e. the standard deviation of the simulated distribution of $\hat{p}^*$) shows that, in general, the true standard error will be underestimated. Hence, in a multi-center setting, analysts should be careful with assuming binomial standard errors for the inference of incidences unless (1) misclassification is assumed to be rare, (2) the incidence of interest is close to 0 and the specificity is close to unity, or (3) both the incidence of interest and the sensitivity are close to 1.

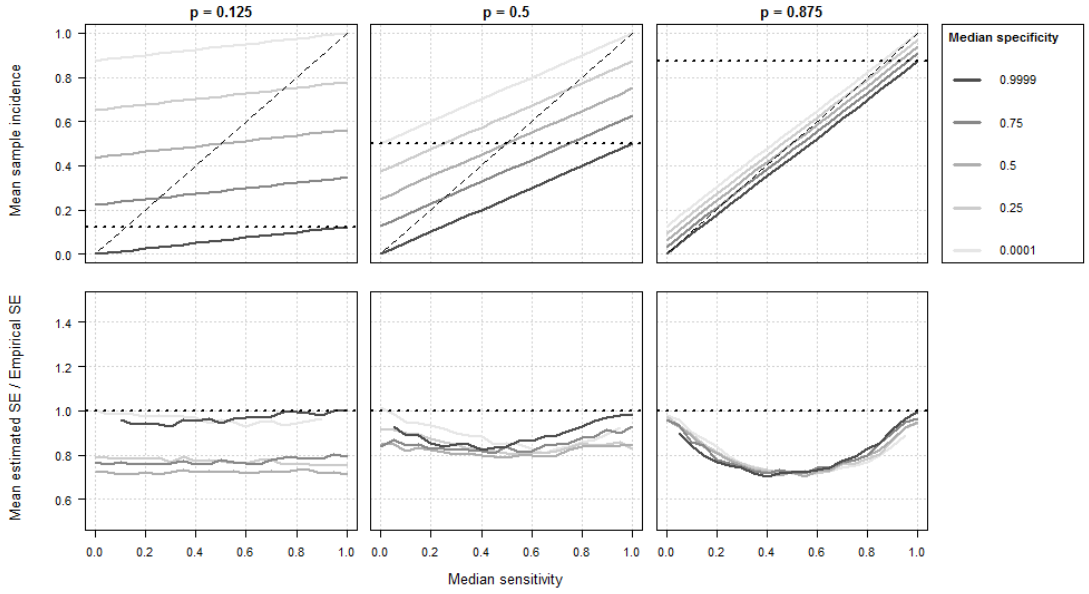

Figure 1. Top: the expected value of $\hat{p}^*$ (i.e. the mean of the simulated estimates) for p equal to 0.125, 0.5, and 0.875 and for a range of median sensitivity and specificity levels in a simulated multi-center study with 5 centers and 50 subjects per center. Dotted horizontal lines indicates the true cumulative incidence probabilities. The dashed line indicates the points on which the median sensitivity and the median specificity sum to one. Bottom: The ratio of the mean value of the naively applied standard estimate of the standard error $\widehat{SE}_{\hat{p}}$ and the empirical standard error (i.e. the standard deviation of the simulated distribution of $\hat{p}^*$).

2.2. Estimation of the treatment effect

To illustrate the impact of misclassification on the estimation of a treatment effect comparing two randomized groups in a clinical trial, let p_0 and p_1 denote the population incidence probabilities for a control and experimental treatment, and again let $\hat{p}_0^*$ and $\hat{p}_1^*$ denote the corresponding apparent sample statistics. Representing the observed trial data as in Table 1, the risk difference is estimated as $\widehat{RD}^* = \hat{p}_1^* - \hat{p}_0^* = \frac{a}{a+b} - \frac{c}{c+d}$, with $\widehat{SE}^*_{\widehat{RD}} = \sqrt{\frac{ab}{(a+b)^3} + \frac{cd}{(c+d)^3}}$. The log risk ratio is estimated as $ln\left(\widehat{RR}^*\right) = ln\left(\frac{a/(a+b)}{c/(c+d)}\right)$ with $\widehat{SE}^*_{ln(\widehat{RR})} =$

$\sqrt{\left(\frac{1}{a}+\frac{1}{c}\right)-\left(\frac{1}{a+b}+\frac{1}{c+d}\right)}$. The log odds ratio is estimated as $ln\left(\widehat{OR}^{*}\right)=ln\left(\frac{ad}{bc}\right)$ with $\widehat{SE}^{*}_{ln(\widehat{OR})}=\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$. The log hazard ratio and its SE are estimated using Cox proportional hazards regression analysis. For the purpose of this paper, we consider an effect statistically significant if the 95% Wald confidence intervals do not contain the null value. Note that the effect estimates are not corrected for center effects. In additional analyses, we apply such a correction by means of regression modelling with a fixed main effect of center (see Appendix B for details).

Treatment	Apparent outcome diagnosis		Total
	Present	Absent	
Experimental	a	b	n_1
Control	c	d	n_0

Table 1. Illustration of a standard fourfold table.

In our simulations, we estimate the effect measures in four illustrative scenarios (Table 2), varying in terms of p_0 (0.25 versus 0.75) and effect size (true risk difference of -0.10 versus 0.05). The number of subjects per center equals 50 in each setting, but the number of centers (m) is adjusted to ensure a power of approximately 0.8 with $\alpha = 0.05$. We assume misclassification to be non-differential.

	Scenario			
	1a	1b	2a	2b
p_0	0.25	0.25	0.75	0.75
p_1	0.15	0.30	0.65	0.80
Required sample size*	496	2496	652	2182
Number of centers (m)	10	50	13	44
Subjects per center	50	50	50	50

*Table 2. Characteristics of the simulated settings.*Minimum sample size requirements per exposure arm were calculated as* $\frac{p_1(1-p_1)+p_0(1-p_0)}{(p_1-p_0)^2}\cdot(1.96+0.84)^2$ *(i.e.* $\alpha = 0.05$ *and* $\beta = 0.20$*). For simplicity, the same sample size is used for each of the effect estimates.*

Figure 2 shows, for each scenario, the expected values of the effect estimators for a range of median sensitivity and specificity levels. Generally, the estimators are unbiased only if both the specificity and the sensitivity equal unity. The only exception to this observation concerns the log RR, which is unbiased as long as the specificity equals 1. For all estimators, the

average effect is nullified when the median sensitivity and specificity values sum to 1. When their sum is lower than 1, the average effect is reversed. For the log OR, the bias appears limited with low incidence probabilities and high median specificity, or with high incidence probabilities and high median sensitivity. For the log HR, the bias appears limited with low incidence probabilities and high median specificity, but bias is still observed when both the incidence probability and the median sensitivity are high.

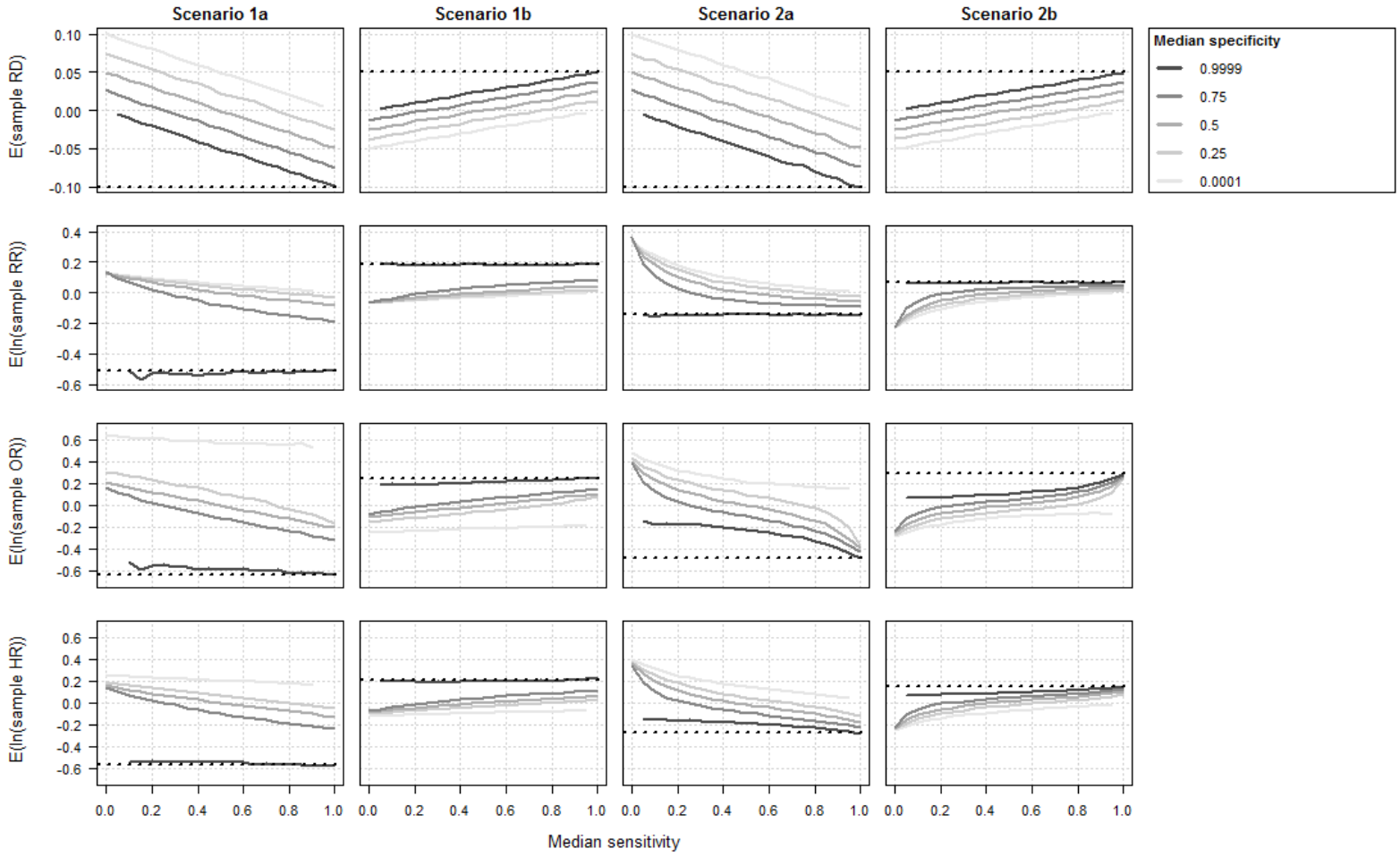

Figure 2. For each simulation scenario, the expected values of the effect estimators (i.e. the risk difference, the log risk ratio, the log odds ratio and the log hazard ratio) for a range of median sensitivity and specificity levels. Dotted horizontal line indicate the true values.

Misclassification also impacts the standard errors, which can be increased or decreased depending on the specific combination of median sensitivity, median specificity, p_0 and p_1 (results not shown). However, the ratio of the average of $\widehat{SE}^*_{\widehat{RD}}$ and the empirical standard error never deviated substantially from 1 (results not shown).

The impact of misclassification on the power of the statistical test is presented in Figure 3. As can be seen, the power decreases rapidly with decreasing sensitivity and specificity, with the relative importance of the median sensitivity being larger in the scenarios with higher incidence probabilities. No substantial differences were observed between the effect measures.

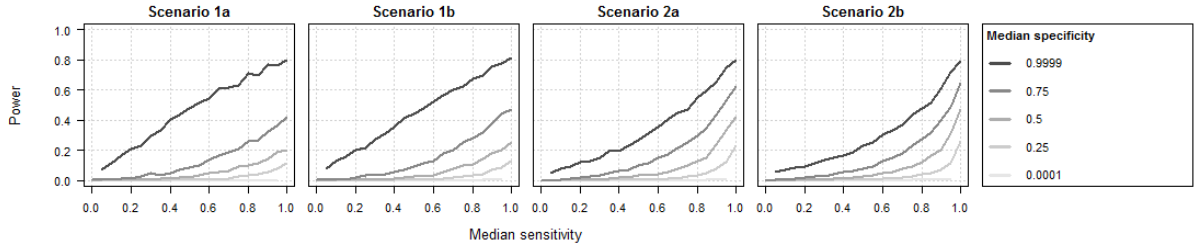

Figure 3. For each simulation scenario, the empirical power of the comparison of incidence probabilities in the two treatment arms. Results shown are based on the analysis of the risk difference, but no apparent differences were observed for the other effect measures. Note: the simulated settings were designed to achieve a power of .8.

3. Effectiveness of adjudication

The illustrations presented in the previous section clarify that misclassification of outcome events in clinical trials (even if non-differential) has the potential to invalidate their results and conclusions. Therefore, the likelihood of misclassification should be minimized. One possible strategy to achieve this is adjudication.

Although adjudication may be an adequate solution, two remarks should be made. First, the term ‘adjudication’ is commonly used specifically for the central assessment of those subjects that experienced the event of interest according to the local investigator, the aim being the detection of false positives. Such approaches may be suboptimal, as detecting missed events, e.g. by means of centralized checks of available data, may be just as important (note that, for the estimation of incidences, a sole focus on false positives may actually introduce bias) [1].

For this reason, we prefer to use the term ‘adjudication’ in the broader sense, including both aims.

Secondly, adjudication is based on the notion that the diagnoses made by the adjudication committee will be more accurate than diagnoses made by investigators. While plausible, various authors have addressed the need for empirical verification of this assumption. For instance, Pogue et al. state that *“the implicit belief that the diagnosis made by a committee is superior and more accurate than an experienced clinician involved in the care of a patient has never been validated”* [3, p. 243]. Ghali [9] makes a similar remark.

In light of these remarks, an assessment of the effectiveness of an adjudication procedure is preferably performed by comparing sensitivity and specificity levels before and after adjudication. However, doing so is typically not feasible due the absence of a gold standard. In practice, therefore, trial-specific pre- and post-adjudication results are compared directly to investigate whether there are substantial differences.

The conclusions of such comparisons vary, with some authors concluding that adjudication has a clear impact (e.g. [10, 11]), while others do not (e.g. [2,4]). However, in their meta-analysis, Pogue et al. [3] were unable to detect, for both masked and unmasked trials, an effect of event adjudication on the estimates of odds ratios for the primary outcomes myocardial infarction, stroke, or cardiovascular/vascular death, or on the number of observed events. A second, more recent meta-analysis also concludes that, on average, odds ratios estimated from adjudicated and non-adjudicated data, collected in trials where investigators are blinded to treatment allocation, did not differ (in this analysis, no explicit comparison of the number of events was performed) [5].

While such comparisons may be useful, their results should be interpreted with care: similar pre- and post-adjudication results may be indicative of a sufficiently low misclassification rate in the initial diagnoses (making the adjudication procedure obsolete), but also of an ineffective adjudication procedure (in which case the post-adjudication data still contain misdiagnosed (non-)events). In addition, such studies tend to focus on differences in the

estimates for a specific type of effect measure (e.g. the sample odds ratio). However, we believe the aim of a trial is more general and includes the estimation of incidence rates and achieving an acceptable level of statistical power. Third, it should be noted that studies in which pre- and post-adjudication results differ may be underreported, because, as stated by Ndounga Diakou et al., *"investigators might be less prone to report the results of both [adjudication committees] and onsite assessors if the results differed"* [5, p.10].

	p_0	p_1	RD	RR	OR	HR	Theoretical power
True values	0.25	0.15	-0.10	0.60	0.53	0.56	0.8
	$E(\hat{p}_0^*)$	$E(\hat{p}_1^*)$	$E(\widehat{RD}^*)$	$E(\widehat{RR}^*)$	$E(\widehat{OR}^*)$	$E(\widehat{HR}^*)$	**Empirical power**
Observed: no adjudication	0.35	0.29	-0.06	0.84	0.77	0.81	0.29
Observed: adjudication (focus on false positives)	0.23	0.15	-0.08	0.68	0.63	0.65	0.57
Observed: adjudication (focus on false positives and false negatives)	0.27	0.18	-0.09	0.67	0.60	0.64	0.69

Table 3. Illustration of the impact of adjudication for simulation scenario 1a on the estimators of the incidences, the effect estimates, and the statistical power. It is assumed that the initial median sensitivity and specificity levels equal 0.8, and that the number of detected false positive and false negative diagnoses is binomially distributed success probability 0.8. Results are based on 2000 simulations. The empirical power is calculated as the proportion of simulations for which the 95% Wald CI for the log(HR) did not contain 0.

As a numerical example, consider again simulation scenario 1a. The true values are given in the top row of Table 3 (note again that the trial was designed to achieve a power of approximately 0.8). Suppose that the initial diagnosis is performed with median sensitivity and median specificity of 0.8. The expected values, simulated under the assumption that no adjudication is performed at all, are provided in the second row. The third row shows the results if adjudication is performed in the traditional sense (i.e. the sole focus being the detection of false positives), and assuming that the average probability that the adjudication committee detects a false positive diagnosis equals 0.8. Lastly, the fourth row provides the result if adjudication focuses on the detection of false positives and false negatives (and again assuming that the probability with which the adjudication committee detects a false negative or false positive diagnosis equals 0.8).

Although it is difficult to determine whether the assumptions made for this illustration are realistic, they do show the possible negative consequences of not performing any adjudication or performing adjudication with a one-sided focus.

4. Discussion

Adjudication of outcome events in clinical trials typically is a resource-intensive procedure, and it is therefore important to justify its need before implementation. In this paper, we showed, by means of an illustrative set of simulated scenarios, that non-differential misclassification can have substantial negative consequences for the validity of trial results, even with realistic sensitivity and specificity levels. These findings contradict the notion that 'random' misclassification errors may not be worthy of correction (note, however, that a formal comparison of our findings and those of Buyse et al. [7], is not straightforward because the authors do not provide the sensitivity and specificity levels of their simulated classification process).

We expected that, on average, structural differences would be observed when comparing pre- and post-adjudication results. However, the results of two meta-analyses appear to suggest that the impact of adjudication may be limited [3, 5]. Although we outlined several possible reasons for this discrepancy, we believe more research is needed to be able to disentangle the problem. For future trials in which adjudication is considered, however, we stress that (1) the impact of misclassification, even if non-differential, should be investigated and not be based solely on the outcome of empirical pre- and post-adjudication results, (2) simulation exercises such as these illustrations may be part of the trial design, based on expected misclassification rates and detection rates of an adjudication committee, in order to get a more detailed insight on the expected gains, and (3) adjudication is one approach to reduce misclassification, but alternatives (e.g. choosing a less subjective outcome, aligning trial outcome definitions with those used in clinical practice, improving protocol training, including only highly trained investigators, and statistical corrections) may be more cost-effective. Lastly, note that we assessed the impact of misclassification and adjudication on the outcomes of clinical trials, but ignored an important potential second purpose: ensuring the safety of individual trial participants.

Appendix A: Simulation details

All simulations were performed in R [12]. The simulation algorithms that were used are detailed below.

Simulation algorithm used for the investigation of the impact of misclassification on the estimation of the cumulative incidence probability

1. Sample center-specific sensitivity and specificity levels θ_j and τ_j on the logit scale from a bivariate normal density with means equal to $logit\left[med\left(\theta_j\right)\right]$ and $logit\left[med\left(\tau_j\right)\right]$ (i.e. the means correspond to pre-specified median values $med\left(\theta_j\right)$ and $med\left(\tau_j\right)$ when transformed back to the original scale) and covariance matrix $\boldsymbol{\Sigma}$. We assume a covariance of 0 and a variance of 0.1 for both variables.
2. For each subject i, determine the true event status y_i as $y_i \sim Binom(p)$.
3. If $y_i = 1$, determine the apparent event status y_i^* as $y_i^* \sim Binom(\theta_j)$. If $y_i = 0$, determine it as $y_i^* \sim Binom(1 - \tau_j)$.
4. If there is at least one apparent non-event and one apparent event, estimate the incidence probability and the corresponding SE.
5. Repeat steps 1-4 5000 times for every investigated combination of $med(\theta_k)$ and $med(\tau_k)$, where $med(\theta_k)$ is evaluated at $(0.0001, 0.05, 0.10, 0.15, \ldots, 0.95, 0.9999)$ and $med(\tau_k)$ at $(0.0001, 0.25, 0.50, 0.75, 0.9999)$.

Simulation algorithm used for the investigation of the impact of misclassification on the estimation of the risk difference, risk ratio, and odds ratio.

1. Sample center-specific sensitivity and specificity levels θ_j and τ_j on the logit scale from a bivariate normal density with means equal to $logit\left[med\left(\theta_j\right)\right]$ and $logit\left[med\left(\tau_j\right)\right]$ and covariance matrix $\boldsymbol{\Sigma}$. We assume a covariance of 0 and a variance of 0.1 for both variables.
2. Per center, randomize subjects in a 1:1 ratio, using block randomization with block size four, to the control treatment arm ($A = 0$) or to the experimental treatment arm ($A = 1$).

3. Simulate y_i as $y_i \sim Binom(p_1)$ if the subject is in arm $A = 1$ and as $y_i \sim Binom(p_0)$ otherwise.
4. If $y_i = 1$, determine the apparent event status y_i^* as $y_i^* \sim Binom(\theta_j)$. If $y_i = 0$, determine it as $y_i^* \sim Binom(1 - \tau_j)$.
5. If there is at least one apparent non-event and one apparent event in both treatment arms, estimate the apparent RD, log RR and log OR and the corresponding SEs. Determine whether the estimates are significantly different from zero in the correct direction.
6. Repeat steps 1-5 2000 times for every investigated combination of $med(\theta_k)$ and $med(\tau_k)$.

Simulation algorithm used for the investigation of the impact of misclassification on the estimation of the hazard ratio.

1. Sample center-specific sensitivity and specificity levels θ_j and τ_j on the logit scale from a bivariate normal density with means equal to $logit[med(\theta_j)]$ and $logit[med(\tau_j)]$ and covariance matrix $\boldsymbol{\Sigma}$. We assume a covariance of 0 and a variance of 0.1 for both variables.
2. Per center, randomize subjects in a 1:1 ratio, using block randomization with block size four, to $A = 0$ or $A = 1$.
3. Assuming that the hazard of outcome event Y is given by $\lambda(Y|A) = \lambda_0 \cdot exp(\lambda_1 A)$, and that observations are censored after 365 days, the cumulative survival function is given by $S(Y|t, A) = exp\left(-\int_0^t \lambda(u)\, du\right) = exp(-t \cdot \lambda_0 \cdot exp(\lambda_1 A))$. $S(Y|t = 365, A = 0) = 1 - p_0$. Therefore, $\lambda_0 = \frac{-ln(1-p_0)}{365}$. Also, because $S(Y|t = 365, A = 1) = 1 - p_1$, $\lambda_1 = ln\left(\frac{\ln(1-p_1)}{\ln(1-p_0)}\right)$. Note that the true HR therefore equals $\frac{\ln(1-p_1)}{\ln(1-p_0)}$. For every subject i, draw $U_i \sim Unif(0,1)$. Simulate true event times as the first $t = (1, \dots, 365)$ for which $U_i > S(Y|t, A)$ minus one half. If U_i never exceeds $S(Y|t, A)$, censor the subject at 365 days.
4. Introduce false negative events: For subjects who experienced an event in center j, censor the true event time at 365 days if a draw from $Binom(\theta_j) = 0$.

5. Introduce false positive events: Within the set of subjects with censored true event times in center j, introduce event times as the first t for which $U_i^{FP} > S(FP|t)$, minus one half. Here, $U_i^{FP} \sim Unif(0,1)$ and $S(FP|t) = exp\left(\frac{t \cdot ln(\tau_j)}{365}\right)$. The latter follows from assuming that the hazard of a false positive event is assumed constant and that $S(FP|t = 365) = \tau_j$. If U_i^{FP} never exceeds $S(FP|t)$, let the event remain censored at 365 days.
6. If there is at least one apparent non-event and one apparent event in both treatment arms, estimate the apparent HR on the log scale and the corresponding SE. Determine whether the estimate is significantly different from zero in the correct direction.
7. Repeat steps 1-6 2000 times for every investigated combination of $med(\theta_k)$ and $med(\tau_k)$.

Note that results are only shown if at most five percent of the simulated trials yield invalid comparisons (e.g. if no (non-)events are observed in one or both of the treatment arms). If the percentage of invalid comparisons is non-zero but lower than five percent, the invalid comparisons are removed from the data before calculation of the outcomes of interest.

Appendix B

To investigate the impact of adjusting for a (fixed) main center effect, we model the RD and log RR using Poisson regression analysis with the identity and log link function and robust standard errors [13]. The adjusted log OR is estimated through logistic regression analysis, with Firth correction to prevent numerical non-convergence caused by separation [14]. The adjusted log HR is estimated by means of Cox proportional hazards regression analysis. The results are shown in Figure B1, and are not substantially different from the unadjusted results, with the exception that the increased model complexity yields more computational problems. A similar result is observed for the statistical power.

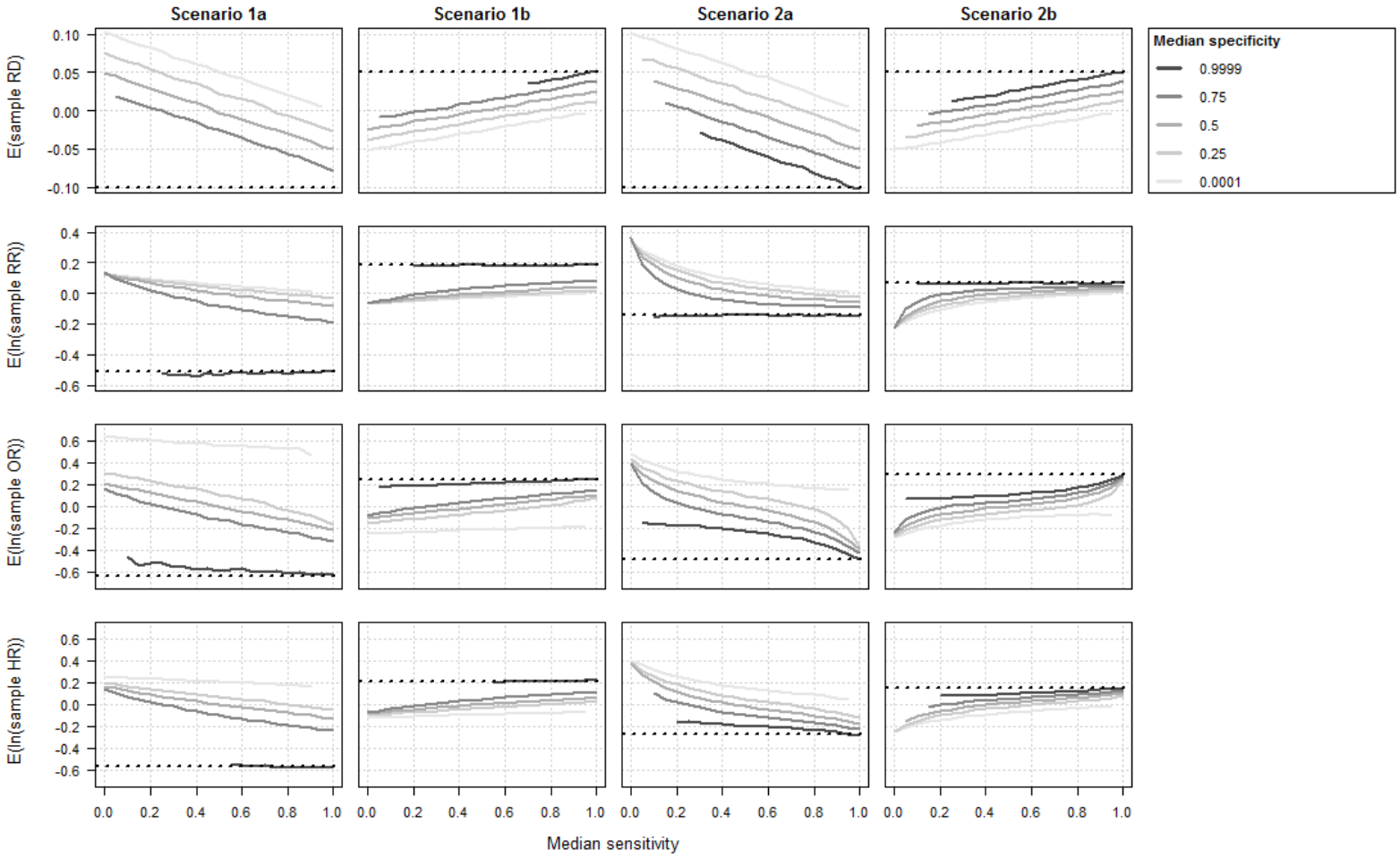

Figure B1. For each simulation scenario, the expected values of the effect estimators (i.e. the risk difference, the log risk ratio, the log odds ratio and the log hazard ratio) for a range of median sensitivity and specificity levels, when adjusting for a center effect. Dotted horizontal line indicate the true values.

References

1. Granger CB, Vogel V, Cummings SR, et al. Do we need to adjudicate major clinical events? Clin Trials 2008; 5:56–60. DOI: 10.1177/1740774507087972.
2. Ninomiya T, Donnan G, Anderson N, et al. for the PROGRESS Collaborative Group. Effects of the end point adjudication process on the results of the Perindopril Protection Against Recurrent Stroke Study (PROGRESS). Stroke 2009; 40:2111-2115. DOI: 10.1161/STROKEAHA.108.539601.
3. Pogue J, Walter SD, Yusuf S. Evaluating the benefit of event adjudication of cardiovascular outcomes in large simple RCTs. Clin Trials 2009; 6:239-251. DOI: 10.1177/1740774509105223.
4. Hata J, Arima H, Zoungas S, et al. on behalf of the ADVANCE Collaborative Group. Effects of the endpoint adjudication process on the results of a randomised controlled trial: The ADVANCE trial. PLoS One 2013; 8(2):e55807. DOI: 10.1371/journal.pone.0055807.
5. Ndounga Diakou LA, Trinquart L, Hróbjartsson A, et al. Comparison of central adjudication of outcomes and onsite outcome assessment on treatment effect estimates. Cochrane Database Syst Rev 2016; 3:MR000043. DOI: 10.1002/14651858.MR000043.pub2.
6. Rothman K, Greenland S, Lash TL. Validity in epidemiologic studies. In Modern epidemiology (third edition). Wolters Kluwer Health, Lippincott Williams & Wilkins 2008.
7. Buyse M, Squifflet P, Coart E, Quinaux E, Punt CJA, Saad ED. The impact of data errors on the outcome of randomized clinical trials. Clin Trials 2017. Published ahead of print. DOI: 10.1177/1740774517716158.
8. Quade D, Lachenbruch PA, Whaley FS, et al. Effects of misclassifications on statistical inferences in epidemiology. Am J Epidemiol 1980; 111(5):503–515.
9. Ghali JK. Do heart failure trials need an end point committee? Am Heart J 2010; 160(4):571–573. DOI: 10.1016/j.ahj.2010.07.005.
10. Näslund U, Grip J, Fischer-Hansen J, et al. The impact of an end-point committee in a large multicentre, randomized, placebo-controlled clinical trial: Results with and

without the end-point committee's final decision on end-points. Eur Heart J 1999; 20:771-777.

11. Mahaffey KW, Harrington RA, Akkerhuis M, et al. for the PURSUIT investigators. Systematic adjudication of myocardial infarction end-points in an international clinical trial. Curr Control Trials in Cardiovasc Med 2001; 2(4):180-186.
12. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org/. Version: 3.3.1.
13. Spiegelman D, Hertzmark E. Easy SAS Calculations for Risk or Prevalence Ratios and Differences. Am J Epidemiol 2005; 162(3):199-200. DOI: 10.1093/aje/kwi188.
14. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med 2002; 21:2409–2419. DOI: 10.1002/sim.1047.

Chapter 6

Collecting and reporting safety data and monitoring trial conduct in pragmatic trials

Elaine Irving
Rutger van den Bor
Paco Welsing
Veronica Walsh
Rafael Alfonso-Cristancho
Catherine Harvey
Nadia Garman
Diederick E Grobbee

on behalf of GetReal Work Package 3

Journal of Clinical Epidemiology, 2017; [Epub ahead of print].

Abstract

Objective: Pragmatic trials offer the opportunity to obtain real-life data on the relative effectiveness and safety of a treatment before or after market authorization. This is the penultimate paper in a series of eight, describing the impact of design choices on the practical implementation of pragmatic trials. *Study design and setting:* This paper focuses on the practical challenges of collecting and reporting safety data and of monitoring trial conduct while maintaining routine clinical care practice. *Conclusion:* Current ICH guidance recommends that all serious adverse events and all drug-related events must be reported in an interventional trial. In line with current guidance, we propose a risk-based approach to the collection of non-drug-related non-serious adverse events and even serious events not related to treatment based on the risk profile of the medicine/class in the patient population of interest. Different options available to support the collection and reporting of safety data while minimizing study-related follow-up visits are discussed. A risk-based approach to monitoring trial conduct is also discussed, highlighting the difference in the balance of risks likely to occur in a pragmatic trial compared to traditional clinical trials and the careful consideration that must be given to the mitigation and management of these risks to maintain routine care.

1. Introduction

Pragmatic trials are usually conducted to demonstrate the real-world effectiveness, safety or health-economic benefits of a new medicine, an existing medicine for a new indication, behavioral/surgical interventions, or diagnostic tests in routine clinical practice [1]. The potential to demonstrate such real-world effects heavily relies on the willingness of patients and treating physicians to participate in this type of research [2, 3] and also on the ability to maintain routine clinical practice (to the extent possible) throughout the duration of the trial. However, this can be challenging due to the regulatory requirements imposed on interventional studies that involve the random assignment of patient to a particular therapy [4]. The collection of adverse events (AE) and the monitoring of trial conduct to ensure that the trial is conducted, recorded, and reported according to good clinical practice (GCP) may impact routine clinical care. This, in turn, would mean that data are not really representative of the real-life clinical situation, and generalizability of trials results may be compromised [5, 6].

However, the nature and extent of AE collection and monitoring of trial conduct can be dependent on the specifics of the study. For example, the concept of "low interventional trials" will be introduced in the new European clinical trial regulation [7]. These regulations relate to investigational medicinal products used according to their marketing authorization or for which the use is evidence based. In addition, the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have stressed the need for developing more risk-based approaches toward the monitoring of trial conduct. These concepts provide an opportunity for a more proportionate approach to the collection of safety data and the monitoring of trial conduct that may be suitable for pragmatic trials. The present paper discusses the main operational challenges in specifying a safety monitoring plan and a trial monitoring plan for pragmatic clinical trials.

1.1. Collection and reporting of safety events

Current international guidelines on GCP state that all serious adverse events (SAEs) must be collected in an interventional trial and reported within 24 hours to the trial sponsor unless specified otherwise in the study protocol [8]. The collection and reporting of non-SAEs can

be tailored under certain conditions, for instance, if a significant amount of safety data is already available as is the case with a marketed medication [9]. Furthermore, the nature and extent of patient safety monitoring may depend on the added risks of the trial intervention relative to standard care [9]. The new European Clinical Trial Regulation moves toward further simplification and defines that the protocol may exclude certain AEs from being recorded and reported and also exclude certain SAEs from requiring immediate reporting [7]. Therefore, here, we propose that the extent to which non-SAEs should be recorded depends on whether (1) the data will be used to define the safety profile of the medication; (2) there are specific safety surveillance measures in place as part of a market authorization; (3) the medication is being used in accordance with the market authorization; (4) the safety profile of the medicine and/or class is known in the patient population under study; and (5) data are being reported by a qualified physician who determines drug relatedness and severity of any given event.

It is conceivable that only serious and non–serious adverse drug reactions (ADRs) may be collected in a pragmatic trial that examines a marketed medicine of which the safety profile has previously been described. However, it should be noted that depending on what is known about the medicine or class of medicines, some drug-related AEs will not be known and therefore careful consideration should be given to assessing the relatedness of AEs with the drug and the risk of missing new drug-related AEs if the decision is made not to report all AEs. In the preauthorization phase or when new indications are investigated, however, collection of all serious and non-SAEs (SAEs and AEs) may still be required (see Figure 1). Furthermore, the higher the risks associated with the medicine and the higher the disease severity, the more frequently physicians should monitor the patients. In pragmatic trials, safety data collection should, wherever possible, be embedded within routine care visits, although it must be acknowledged that these vary widely globally according to different local practice. The frequency with which data should be collected will also depend on the level of safety information available about the medicine. The less information available or the higher the risks associated with the medicine, the more regularly physicians should monitor the patients. In cases where local norms are insufficient, visit frequency may need to be included in a protocol. The disease being studied, and trial design will also affect the frequency of data

collection. More severe conditions, which are normally associated with more frequent routine monitoring by the treating physician, offer greater opportunity to also frequently collect study-related safety data in a manner that does not interfere with routine care.

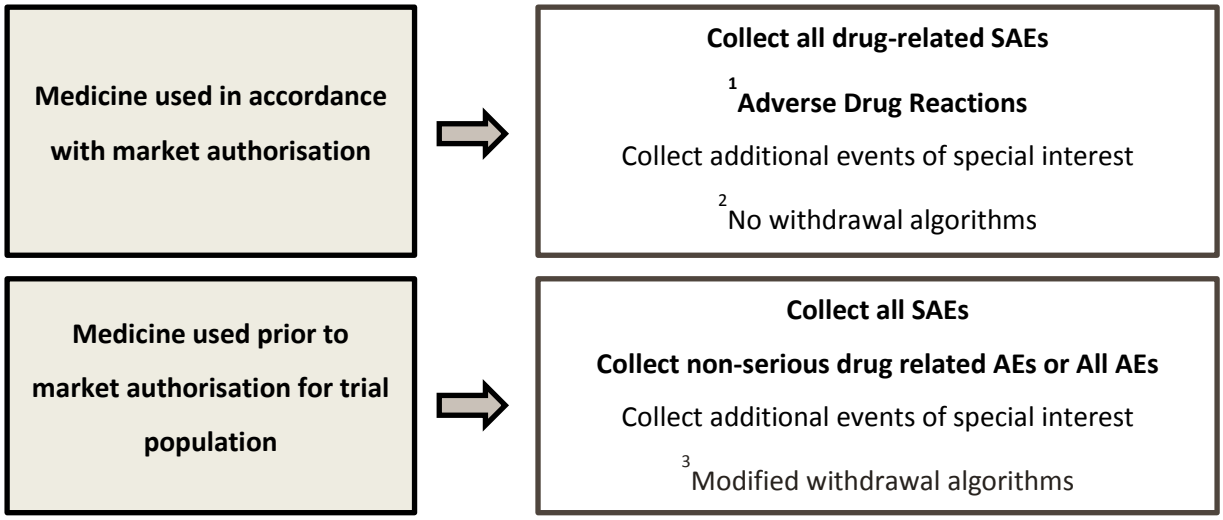

Figure 1. Which safety events should be collected in a pragmatic trial? [1]Level of collection will depend on: objectives of the study effectiveness or effectiveness/risk, what is known about the safety profile of the drug, postmarketing pharmacovigilance requirements, and length of time drug has been on the market; [2]Patients should be managed in accordance with standard clinical practice; there is no need to impose additional "trial" restrictions; [3]Modified algorithms to accommodate comorbidities in trial population, Serious adverse event (SAE): any serious medical occurrence in a clinical study participant, temporally associated with the use of a study treatment, whether or not considered related to the study treatment.

1.2. The challenges of collecting and reporting AE data in pragmatic trials

As pragmatic trials investigate the effects of medication in routine care, they may include investigators who are inexperienced in conducting clinical research, AE reporting, and assessing drug relatedness. From a practical perspective of costs, time, and space, frequent trial visits may not be something the practice can easily manage and so may affect the willingness of physicians and patients to participate. Some severe conditions, such as Chronic obstructive pulmonary disease (COPD), require a high frequency of follow-up visits; however, the frequency often decreases as the control of symptoms improves, and patients may not be seen for several months. Similarly, patients with well-controlled high blood pressure may only visit their physician annually for routine check-ups, introducing the need to consider alternative methods, such as follow-up by telephone, to collect safety data. Furthermore, study patients may present themselves in a health care department or emergency room outside the scope of the investigator. This and any associated safety events may remain unreported to the patient's physician, especially with infrequent follow-up, unless it is

recorded in the hospital medical records or the investigator is reminded to interview the patient on clinical events during a follow-up.

1.3. Possible approaches to the collection and reporting of AEs

Collecting safety data in a structured and complete manner is crucial to understand the course of an AE and its potential drug relatedness. The mechanism for collecting and reporting information about an AE, as well as information about the relevant context such as comedication/comorbidities, will depend on (1) the objectives and design of the trial; (2) whether the patient is randomized to any treatment and whether those treatments are already licensed or not; (3) the normal treatment pathway of the patient in the country where the trial is being conducted (i.e., the regularity with which patients will routinely visit their treating physician or whether the patient routinely visits multiple physicians for different aspects of their health care); and (4) the availability of electronic health record (EHR) data or the use of case report forms for data collection and reporting. A variety of options are described below and summarized in Figure 2.

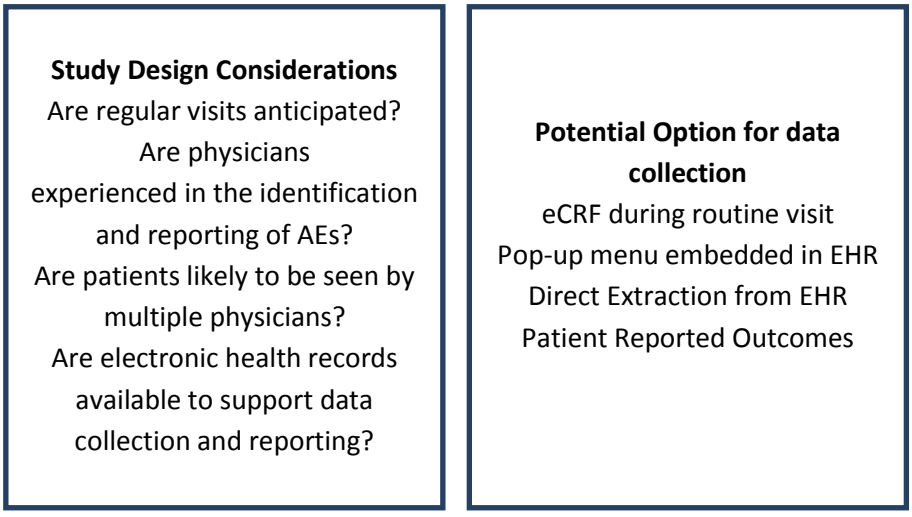

Figure 2. Potential options for data collection. EHR: electronic health record.

When thinking about how to gather safety data in the absence of trial visits, where a marketed product is being used, it may be attractive to consider the use of the spontaneous reporting systems that are in place in many countries. These systems have been put in place to enable patients and physicians to report AEs that they believe are associated with a drug [10, 11]

directly to their local authorities. The advantage of this type of approach is that the market authorization holder can continually monitor the systems for any reported events as part of their standard pharmacovigilance activities. However, spontaneous reporting systems are not designed to and do not support reporting for a trial as the source of data cannot be traced, and information of the event is often incomplete. This may lead to problems with assigning relatedness to the medicines. Often, there is no scope within the systems to (re-)evaluate the relatedness of ADRs submitted by patients or physicians to drug exposure or to specifically label potential events as AEs instead of ADRs. Furthermore, there is known underreporting of events for medicines with which patients and physicians are familiar. In general terms, therefore, spontaneous reporting systems support the detection of potential unknown ADRs; however, they are not suited for estimating the incidence of ADRs [10, 12].

Patient-reported outcomes collected both via validated instruments and web-based sources are becoming increasingly important in understanding the benefits and risks of medicines in the real world. Indeed, numerous pharmacovigilance programs are already integrating structured patient reporting, with evidence of improved detection of treatment risks [13]. However, to be accepted by industry and regulators, there needs to be consensus on what and how data should be collected; how safety signals should be managed; and whether or not physician confirmation is appropriate when patient or carers are reporters.

EHRs are adopted by an increasing number of practices and hospitals and can therefore be instrumental in collection of information about specific AEs. To facilitate a more structured enquiry about symptoms, events, or other relevant information, electronic pop-up menus incorporated in the EHR system could be helpful and indeed have been shown to support accurate reporting in pragmatic trials [6]. However, given the fact that EHR systems are developed to support patient care rather than clinical research, many systems are not validated for use in clinical research or meet the requirements of international safety reporting [14, 15]. If EHR systems are to be used, it is important to discuss and understand how physicians enter information about potential AEs into the EHR system. Known ADRs are unlikely to be the sole reason for a patient to visit their physician, any detailed information about the event may only be captured as free text, and therefore, it is essential to ensure that the free text fields of

the EHR system can be searched for information that is relevant for AE reporting to limit under reporting of these ADRs. On the other hand, an event that was not already known to be associated with the medication may actually be a reason for consulting the physician but may not be labeled as such. The level of missing data, misclassification of events, and potential lag between an event happening and the availability of data for expedited reporting are all aspects that will impact the suitability of EHRs systems for reporting events. For example, to depend on EHR surveillance fully, the EHR system needs to capture all physician encounters both within the trial setting and also any other settings where the patient may receive medical care. It must also capture all necessary data to support the reporting and interpretation of the event, such as the time to onset of event, a description of the event, seriousness criteria, drug relatedness, and the outcome of the event as well as drugs prescribed at time of event and information regarding related comorbidities.

The Salford Lung Studies have been designed to evaluate the effectiveness and safety of the once-daily combination of the ICS fluticasone furoate and the novel LABA vilanterol (VI) (Relvar) compared with existing maintenance therapy in a large, real-world population of patients with COPD [16] and asthma [17] in conditions of normal care. Both these studies were supported by an integrated primary and secondary EHR system which enabled physicians to monitor the occurrence of potential AEs in near real time. Data were reviewed daily by an independent trial safety team of physicians and nurses. If events were identified, the trial team contacted the treating physician to establish the relatedness of the event to the drug [16, 17]. Such integrated systems offer the potential for near real-time safety reporting to the trial sponsor, which is comparable or even beyond that currently available in the routine clinical trial setting where non–serious events are generally collected at least every 4 weeks.

1.4. Monitoring of trial conduct

Regulatory agencies have stressed the need for pharmaceutical companies to develop more proportionate and risk-based methodologies for clinical trial monitoring [18, 19], as the resource-intensive “traditional” approach to trial monitoring becomes less feasible as trial complexity increases. This point of view corresponds to the current ICH GCP guidelines,

which explicitly dictate that *"the sponsor should develop a systematic, prioritized, risk-based approach to monitoring clinical trials"* [20, p.30].

Various risk assessment tools have been developed that can aid protocol developers in establishing the risk level of a trial by considering the wide range of possible sources of issues that may occur during the conduct of clinical trials. For example the procedures used in the ADAMON and OPTIMON studies [21, 22] or TransCelerate's Risk Assessment Categorization Tool [23]. These tools may be useful for pragmatic trials as well, provided it is taken into account that the likelihood of risks occurring in a pragmatic trial or the impact may be different from a traditional trial. For example, failure to follow the protocol is often cited as being one of the top five deficiencies of efficacy trials when audits are conducted by regulatory authorities. In contrast, in a pragmatic trial, because the protocol should remain very flexible to reflect standard clinical practice, less protocol deviations may be expected. If pragmatic trial protocols request data collection that is not part of routine care but regarded as of key importance to the trial, the risk of missing data (which may be regarded a protocol deviation) may be higher in pragmatic than in traditional randomized controlled trials due to the generally less strict follow-up requirements and the participation of a higher proportion of research-naive physicians. Treatment adherence is likely to be much lower in a pragmatic trial and would be considered a risk in a traditional trial. However, lack of adherence may actually drive the real-world effectiveness or safety of the medicine and is therefore not a risk that would be actively managed in a pragmatic trial.

The main challenge when designing a monitoring plan for a pragmatic trial is defining how to manage each of these risks. For example, in a traditional trial, a site visit and retraining may be triggered if investigators continually fail to accurately complete the electronic case report form (eCRF) and/or report SAEs in a timely manner. In the case of a pragmatic trial, especially where data are being extracted directly from the EHR, some missing data are inevitable and should be factored into the protocol design and data analysis plans. Although large problems with missing data may trigger some (re-)training, one would be reticent with this as it may interfere with the real-world nature of the study. The use of broader inclusion criteria, as is typical for pragmatic trials, may in fact result in less favorable risk assessments,

as more vulnerable patients are included. The same may be true when including investigators that are less experienced. Each of these could trigger the requirement for additional site visits which may impact on routine care.

Various approaches to monitoring are being implemented, and central monitoring approaches are being more routinely adopted [24]. For instance, statistical monitoring may help in identifying unlikely or inconsistent data patterns (and may in that respect be preferred over on-site visits) [25]. Random and targeted source data verification (SDV) may reduce the volume of data that is to be SDV-ed by the clinical research assistant (CRA) [26-29]. Remote SDV may remove the need to visit trial sites physically [30, 31]. The use of centrally available data allows for a dashboard-like overview of key performance indicators or metrics, which may be used to trigger on-site visits if specific thresholds are exceeded (see [23] for examples). However, adopting centralized forms of monitoring that require regularly updated sources of data may be challenging in a pragmatic trial. Often investigators need to apply to access EHR records. Therefore, what data are required and how often it is required will in general have to be decided in the protocol design/site feasibility stage, to ensure access can be sought to allow trial recruitment, data completeness, and other trial performance metrics to be assessed.

In the case of on-site monitoring, it is important to ensure that the EHR systems can support requirement for monitoring with regard to controlled access (e.g., monitors should be given unique credentials to access only those records of patients enrolled in the study, and they should not be able to access other patients records who have not consented to be part of the study). As described above, this is often a challenge when working with EHR systems that have been developed to support patient care rather than clinical research.

In conclusion, the risk-based monitoring framework provides flexibility, which may benefit pragmatic trials, but also requires careful and early consideration of the monitoring plan for each individual trial which may be challenging.

2. Discussion and conclusion

ICH Guidelines, for all clinical trials in any phase, dictate that all SAEs must be reported to the sponsor and within 24 hours of the investigator becoming aware of an event unless directed otherwise in the protocol. In pragmatic trials, that generally follow routine clinical practice, additional study-related activities and follow-up visits should be minimized. This may complicate the structured collection of drug-related safety data. However, several different options for data collection are available to researchers depending on the study design which range from more traditional collection of safety data using an eCRF during routine care visits to the remote surveillance of EHR data.

The risk-based monitoring approaches being used to increase efficiency of monitoring trial conduct can be applied to pragmatic trials. However, although the types of risks associated with the implementation of a pragmatic trial are similar, their relative importance, detail, and/or interpretation are likely to be different from a traditional trial. Remote monitoring techniques currently rely on data being fed through companies' data management systems; this works well where data collection is done via eCRF, however, where data are collected directly from an EHR system a mechanism will need to be put in place to allow monitors to access the data required. This requires up front planning and discussion with data system custodians as it is possible that investigators may need to apply for permission to access the data (even remotely) at set intervals throughout the study.

Whatever approach is taken for collecting safety data and monitoring trial conduct, in a pragmatic trial, it is likely to be “nonstandard” compared to an explanatory trial (Figure 3) and will therefore require early and detailed discussions with ethics review boards and regulatory authorities in all participating countries.

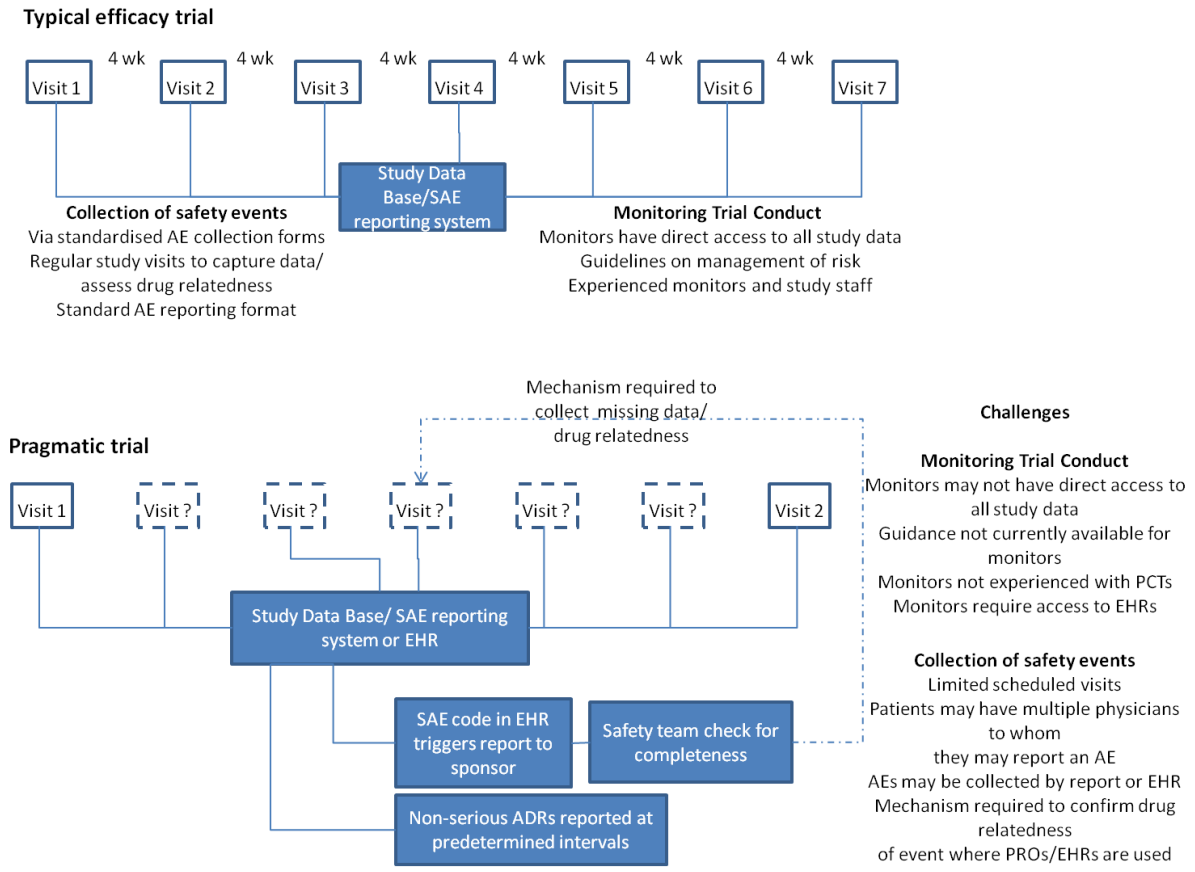

Figure 3. Comparison between efficacy/safety and pragmatic trials. AE: adverse event; SAE: serious adverse event; EHR: electronic health record; ADR: adverse drug reaction; PCT: Pragmatic Clinical Trial; PRO: patient-reported outcome.

References

1. Zuidgeest M, Goetz I, Groenwold RHH, Irving EA, van Thiel G, Grobbee DE. Pragmatic trials and real-world evidence. J Clin Epidemiol 2017; 88:7-13.
2. Rengerink KO, Kalkman S, Collier S, Ciaglia A, Worsley S, Lightbourne A, Eckert L, Groenwold RHH, Grobbee DE, Irving EA. Participant eligibility, recruitment and retention in pragmatic trials. J Clin Epidemiol 2017; [Epub ahead of print].
3. Worsley SD, Rengerink KO, Irving EA , Lejeune S, Mol K, Collier S, Groenwold RHH, Enters-Weijnen C, Egger M, Rhodes T. Setting, sites, and investigator selection. J Clin Epidemiol 2017; 88:14-20.
4. European Trial Directive 2001/20/EC. URL: http://ec.europa.eu/health/files/eudralex/vol-1/dir_2001_20/dir_2001_20_en.pdf
5. Konstantinou GN. Pragmatic trials: how to adjust for the 'Hawthorne effect'? Thorax 2012; 67:562.
6. van Staa TP, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B, Cassell J, Pirmohamed M, Torgerson D, Ronaldson S, Adamson J, Taweel A, Delaney B, Mahmood S, Baracaia S, Round T, Fox R, Hunter T, Gulliford M, Smeeth L. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. Health Technol Assess 2014; 18:1-146.
7. Regulation (EU) No 536/2014 of The European Parliament and of the Council. URL: http://ec.europa.eu/health/files/eudralex/vol-1/reg_2014_536/reg_2014_536_en.pdf
8. Guideline for good clinical practice E6(R2). 2015. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/08/WC500191488.pdf
9. MRC/DH/MHRA Joint Project. Risk-adapted Approaches to the Management of Clinical Trials of Investigational Medicinal Products. URL: http://webarchive.nationalarchives.gov.uk/20150111011944/http://www.mhra.gov.uk/home/groups/l-ctu/documents/websiteresources/con111784.pdf
10. McLernon DJ, Bond CM, Hannaford PC, Watson MC, Lee AJ, Hazell L, Avery A; Yellow Card Collaboration. Adverse drug reaction reporting in the UK: a

retrospective observational comparison of yellow card reports submitted by patients and healthcare professionals. Drug Safety 2010; 33:775-88.

11. de Langen J, van Hunsel F, Passier A, de Jong-van den Berg L, van Grootheest K. Adverse drug reaction reporting by patients in the Netherlands: three years of experience. Drug Safety 2008; 31:515-524.
12. Hoffman KB, Dimbil M, Erdman CB,Tatonetti NP, Overstreet BM. The Weber Effect and the United States Food and Drug Administration's Adverse Event Reporting System (FAERS): Analysis of Sixty-Two Drugs Approved from 2006 to 2010. Drug Safety 2014; 37:283-294.
13. Banerjee A, Ingate S. Web-based patient-reported outcomes in drug safety and risk management: challenges and opportunities? Drug Safety 2012; 35:437-446.
14. Electronic Health Records MHRA Position Statement. URL: http://forums.mhra.gov.uk/showthread.php?1885-Electronic-Health-Records-MHRA-Position-Statement
15. Meinecke A-K, Welsing P, Kafatos G, Burke D, Trelle S, Kubin M, Nachbaur G, Egger M, Zuidgeest M. Data Collection in pragmatic trials. J Clin Epidemiol 2017; [Epub ahead of print].
16. Bakerly ND, Woodcock A, New JP, Gibson JM, Wu W, Leather D, Vestbo J. The Salford Lung Study protocol: a pragmatic, randomised phase III real-world effectiveness trial in chronic obstructive pulmonary disease. Respir Res 2015; 16:101.
17. Woodcock A, Bakerly ND, New JP, Gibson JM, Wu W, Vestbo J, Leather D. The Salford Lung Study protocol: a pragmatic, randomised phase III real-world effectiveness trial in asthma. BMC Pulm Med. 2015; 15:160.
18. FDA. Oversight of Clinical Investigations - A Risk-Based Approach to Monitoring. 2013. URL: http://www.fda.gov/downloads/Drugs/.../Guidances/UCM269919.pdf
19. EMA. Reflection paper on risk based quality management in clinical trials. 2013. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/11/WC500155491.pdf
20. ICH GCP. Integrated addendum to ICH E6(R1): guideline for good clinical practice E6(R2). URL:

http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R2__Addendum_Step2.pdf

21. Brosteanu O, Houben P, Ihrig K, Ohmann C, Paulus U, Pfistner B, Schwarz G, Strenge-Hesse A, Zettelmeyer U. Risk analysis and risk adapted on-site monitoring in noncommercial clinical trials. Clin Trials 2009; 6:585-596.
22. Chêne G, Alberti C, Bellissant E, Bénichou J, Carrat F, Chatellier G, Costagliola D, Demotes-Mainard J, Guillemin F, Leizorovicz A, Pignon JP, Preux PM, Rascol O, Ravaud P, Tréluyer JM. Evaluation of the efficacy and cost of two monitoring strategies for public clinical research. OPTIMON trial: OPTImisation of MONitoring of clinical research studies (protocol). 2008. URL: https://ssl2.isped.u-bordeaux2.fr/OPTIMON/Documents.aspx
23. TransCelerate: Position Paper Risk Based monitoring. 2013. URL: http://www.transceleratebiopharmainc.com/wp-content/uploads/2013/10/TransCelerate-RBM-Position-Paper-FINAL-30MAY2013.pdf
24. Morrison BW, Cochran CJ, White JG, Harley J, Kleppinger CF, Liu A, Mitchel JT, Nickerson DF, Zacharias CR, Kramer JM, Neaton JD. Monitoring the quality of conduct of clinical trials: a survey of current practices. Clin Trials 2011; 8:342-349.
25. Baigent C, Harrell FE, Buyse M, Emberson JR and Altman DG. Ensuring trial validity by data quality assurance and diversification of monitoring methods. Clin Trials 2008; 5:49-55.
26. Getz K. Low hanging fruit in the fight against inefficiency: Direction from regulatory agencies would help eradicate wasteful 100 percent source data verification. Appl Clin Trials, March 2011.
27. Hines S. Targeting Source Document Verification: Targeted SDV provides data validation by comparing trial data with primary health records. Applied Clinical Trials, February 2011.
28. Grieve AP. Source Data Verification by Statistical Sampling: Issues in Implementation. Drug Inf J 2012; 46:368-377.

29. Tantsyura V, Grimes I, Mitchel J, Fendt K, Sirichenko S, Waters J, Crowe J, Tardiff B. Risk-based source data verification approaches: pros and cons. Drug Inf J 2010; 44:745-756.
30. Radovich C, Frick J. Remote Source Document Verification (rSDV): A Sponsor Perspective and Results of Implementation. Monitor 2009. URL: http://www.pharmavigilant.com/uploadDocs/1/December-Monitor-article-on-rSDV.pdf
31. Uren SC, Kirkman MB, Dalton BS, Zalcberg JR. Reducing clinical trial monitoring resource allocation and costs through remote access to electronic medical records. J Oncol Pract 2013; e13-e16.

Chapter 7

General discussion

1. Introduction

The quality of a clinical trial is commonly interpreted as its ability to *"effectively answer the intended question about the benefits and risks of a medical product (therapeutic or diagnostic) or procedure, while assuring protection of human subjects."* (see, e.g., [1, p.S112] and [2, p.125]).

To assure the quality of the trial during its conduct phase, different parties perform a range of monitoring procedures [3]. Typically, (1) the health status of individual subjects is monitored by the investigators (i.e., the treating physicians or nurses); (2) the adherence of investigators (and their teams) to the trial protocol is monitored by a Clinical Research Associate (CRA) employed by the sponsor or Contract Research Organization (CRO); (3) the correctness of the data that is being submitted by investigators (and their teams) is being reviewed by data management departments of the sponsor or CRO. In addition, if the assessment of clinical events that are important for the trial requires some level of subjectivity, an adjudication committee may be appointed to review the initial diagnoses; (4) the day-to-day progress of the trial is monitored by the trial management committee consisting of representatives of the sponsor (sometimes in combination with other experts, in which case the committee is called the 'trial steering committee'); (5) the emerging risk/benefit profile of the investigational medicinal product and the scientific validity of a trial are monitored by institutional review boards and the data monitoring committee.

As trials are becoming ever more complex and costly [4-6], it is necessary to critically reflect on how resources are being spent. In this respect, trial monitoring activities, in particular those related to the monitoring of investigators and their teams by CRAs, have frequently been the subject of critical remarks.

Typical tasks of a CRA include the training of site personnel, validating informed consent forms, inspection of unreported adverse events, and checking for errors in the data [7, section 5.18.4]. Traditionally, with trials being largely paper-based, site monitoring was largely performed at the investigational center, requiring the CRA to physically visit the center on regular intervals during the progress of the trial. One of the most time-consuming activities of

the CRA is source data verification (SDV). SDV is the process of comparing source data against the data as written/entered on (electronic) case report forms to determine whether information was copied incorrectly (or not at all) and to correct any discrepancies. In industry-sponsored trials in particular, frequent site visits in combination with SDV of all source data ('100% SDV') has long been the standard ([4,8]).

This intensive 'standard' approach towards site monitoring has been the subject of much debate, with criticism focused on (1) its substantial costs [9-14], (2) the lack of empirical evidence favoring such methods over possibly more efficient alternatives [4,9,12,15-17], (3) the lack of proportionality, i.e., the implicit view that all errors are equally important [9, 18], and (4) the limited scope of SDV, which aims to detect transcription errors but may fail to detect inaccuracies in the source documentation [19]. In addition, from a quality management perspective, it can be argued that *"the current system of inspecting clinical sites to ensure quality in clinical trials also encourages an approach similar to old-fashioned manufacturing systems: produce the product, catch the defective ones, and throw them out. Throwing out clinical trial data after the fact is ineffective and wasteful."* [1, p.S112] (similar statements are found in [9] and [20]). Lastly, it has been argued that the reliance on frequent site visits with 100% SDV is based on an overly strict interpretation of regulatory guidelines [21,22] (although those guidelines themselves have been subjected to criticism as well [19,23]).

In response, both the European Medicines Agency (EMA) and the Food and Drug Administration (FDA) finalized a guidance report in which they stress the need to adapt more efficient, risk-proportionate approaches towards clinical trial monitoring [4,18]. However, the agencies also acknowledge that *"this is an evolving area"* [18, p.15] and that *"there are limited empirical data to support the utility of the various methods employed to monitor clinical investigations (e.g., superiority of one method versus another), including data to support on-site monitoring"* [4, p.6].

In summary, the discussion on this topic has been going on for many years (e.g. see [9]). Nevertheless, in 2013, there still was limited empirical evidence concerning the effectiveness of the 'traditional' monitoring model (i.e., 100% SDV in combination with frequent site

visits). Therefore, the following questions need to be addressed: from the available literature, can we indeed conclude that the traditional model of site monitoring is insufficiently effective to justify its costs? Which alternatives have been proposed? And how is their impact on the trial quality and costs assessed?

2. The value of SDV

Sheetz et al. [24] assessed the value of SDV by various means (literature review, retrospective analysis of trial data and an assessment of major and critical findings from internal audits) and conclude that *"SDV identifies an insignificant number of transcription errors"* and that *"SDV should no longer be the foremost quality management method employed in clinical trials, as SDV does not meaningfully contribute to overall data quality."* [24, p.679]. The authors also tried to investigate the impact of SDV on safety (in terms of missed (serious) adverse events), but were unable to provide conclusive results. They do note that there may be value of on-site monitoring visits in the detection of unreported adverse events. In a more recent literature review, Olsen et al. [25] conclude that *"[...] 100% SDV is not a rational method of ensuring data integrity and subject safety based on the high cost [...]"* [25, p.411]. Indeed, these results underscore the earlier critical remarks and stress the need for alternative methods that are less dependent on SDV.

3. Proposed alternative site monitoring strategies

As a result of the discussion, various alternative procedures have been proposed. In certain cases, these consist of 'simple' adaptations. For instance, pilot studies show that performing SDV or checking informed consent forms remotely (as opposed to on-site) may be feasible alternatives [26-29]. In addition, it may be attractive to focus SDV on critical aspects only or to use statistical sampling methodology [8,21,30-32].

However, most commonly, proposed strategies are based on the view that the type and nature of monitoring should be made dependent on the outcomes of some type of risk assessment. Generally, such methods are referred to as 'Risk-Based Monitoring' (RBM), although there exists considerable variation (see e.g. [33]). To avoid confusion, we make a rough distinction into three 'types' of RBM strategies (although they may be used in combination):

In its simplest form, a risk assessment is a structured a priori assessment of the risks based on the context and protocol of the trial (see e.g. the ADAMON risk analysis form [34] or TransCelerate's Risk Assessment Categorization Tool [35]). Such assessments may consist of a series of items focusing on patient safety (e.g. 'will the trial pose greater risk than standard treatment?') or on the scientific validity (e.g. 'does the diagnosis of the primary endpoint require a subjective assessment?'). The aim is two-fold: to ensure that the protocol is optimized (and adjusted when possible), and to assure that risks identified are mitigated by the monitoring activities described in the monitoring plan. An a priori assessment of risks can provide useful information regarding the prioritization of risks and the type of information that is necessary to ensure quality throughout the progress of the trial so that problems, when they occur, are quickly detected and mitigated: As reported by one monitor involved in the conduct of this type of RBM, *"the positive effect clearly is that you think more about the study itself. If you take the effort to classify the study by risk factors you can actually eliminate many things upfront. Because of that evaluation, I know what to set value on when I open a site."* [36, p.7].

A second interpretation of RBM concerns the assessment of (regularly updated) information indicative of 'site performance'. 'Site performance' is typically categorized in specific domains such as recruitment/retention, patient safety, protocol adherence, etc., each of which is quantified by one or more so-called 'risk indicators' or 'performance indicators' (see, e.g., [37,38]). The CRA can use this information (which is preferably obtained centrally) to determine which actions are required to ensure that centers are performing sufficiently well. Predefined triggers may be set to automate this process whenever some threshold is exceeded. These thresholds can be fixed values, values relative to the performance of other centers, or values indicative of a 'sudden' worsening.

Besides maintaining oversight on centers through performance indicators, the structured set-up of clinical trials allows between-center comparisons, to search for data pattern deviations indicative of data entry errors, non-calibrated measurement equipment, underreporting of clinical event/endpoints, or data fabrication. Checks are preferably performed in parallel to

the monitoring of risk indicators, i.e. during the conduct of the trial, to detect deviating centers as early as possible. This form of monitoring (often referred to as 'central statistical monitoring' (CSM)) typically consists of graphical or statistical assessments (see, e.g., [39-49]). It is substantially different from the assessment of performance indicators [14], and may therefore require a separate role of 'statistical monitor'. Often, the assessments focus on detecting data that is 'too perfect', or on deviations that can be expected when data is being fabricated (such as deviating patterns of digits [41]). If centers are flagged, further investigation of the data may be performed (potentially resulting in a for-cause audit).

Many examples of the above procedures are publicly available. For instance, the 'RBM toolbox' from the European Clinical Research Infrastructure Network (ECRIN), available on-line [50], provides a list of available tools. Examples of risk indicators are available from the TransCelerate [37] website. Programs that can be used for statistical monitoring are available as well (e.g. [42,49]).

Nonetheless, there remains a need to critically reflect on available RBM before implementing them in practice. In our view, the publications by the EMA and the FDA [4,18] present an opportunity for a careful reassessment of the value and efficiency of the different options a CRA has when monitoring sites, including the utilization of technological advancements (e.g. the availability of trial data in real-time). Importantly, both publications suggest a greater focus on preventing issues from occurring, as opposed to detecting and correcting them.

However, as pointed out by Ansmann et al. [51], proposed alternative monitoring strategies *"almost always focus just on SDV and a reduction in the number of visits"* [51, p. 2], which may be a too narrow focus. For instance, a priori risk assessments commonly yield a single risk category for a given trial. A corresponding visit schedule and 'percentage of SDV' is then assigned. Yet, while the main purpose of SDV is to improve the validity of the trial data by correcting transcription errors, it is not uncommon that the risk assessments are primarily focused around 'risk' in terms of patient safety. In our view, it is unclear whether it is reasonable to accept more transcription errors in studies that pose a low safety risk: the data needs to be of sufficient quality regardless of the exact nature of the investigational medicinal

product or the patient population. We therefore advise to carefully determine whether the relation between the outcome of a certain risk assessment and the proposed mitigation strategy is, in fact, reasonable[6]. In this respect, it may help to avoid assigning a single risk score to a trial, but instead focus on identified risks on a case-by-case basis.

A second consideration relates to the expected reduction in terms of monitoring costs when shifting from on-site monitoring to centralized monitoring. For smaller trials in particular, the adaptation may imply a shift of resources rather than actual cost savings, as the implementation of the required infrastructure and methodology bring costs as well.

4. Assessment of costs and effectiveness

When choosing to alter site monitoring practices, it is necessary to investigate whether the alterations indeed have the desired effect. To some extent, this may be assessed a priori via simulation or retrospective analysis. Yet, in most cases, actual implementation in running trials will be required before conclusive statements can be made. Whenever feasible, such assessments are preferably performed with as much care as any investigation that aims to draw causal inferences, i.e. by means of a prospective study in which investigational centers are randomized to one of two monitoring strategies. At the very least, outcomes (e.g. monitoring costs or the number of important events not detected) should be pre-defined and the design should allow a sensible comparison. Although doing so may require additional resources (such as independent monitoring to observe whether certain events were not detected by the CRA), they are of key importance to ensure that the quality indeed is not negatively affected. Regardless of the exact type of analysis that is performed, it is important to share the results of these investigations in a manner that is transparent and that contains sufficient level of detail concerning the methodology applied. Preferably, via publications in peer-reviewed scientific journals (as opposed to e.g. white papers, which offer limited room for methodological details).

[6]Also note that the relation between anticipated risks and the quality may not be as expected: *"Counterintuitively, high risk studies [...] seem to be at lower risk for poor quality, probably due to the more closely monitored regulatory and legal environment which supports a well-planned set up of the study."* [36, p.11].

There are several studies which are valuable in this respect: First, Liénard et al. [15] investigated the impact of on-site initiation visits on a number of outcomes (patient recruitment, quantity and quality of data submitted, and patient's follow-up time). Centers were randomized to receive on-site initiation visits or not. The authors could not detect a significant contribution of on-site initiation visits to the trial. Second, the ADAMON trial [52] compared 'extensive on-site monitoring' (including complete SDV) against a monitoring strategy adapted to trial-specific risks. The outcome of interest was the number of patients with at least one major violation of Good Clinical Practice objectives. Centers were randomized to one of the monitoring strategies. The authors conclude that the risk-adapted monitoring strategy, while requiring less than half of the monitoring resources, is non-inferior to the extensive monitoring strategy, but also that the overall violation rates were much higher than anticipated. OPTIMON [53], a third study, is comparable to the ADAMON study. First results [54] again show a high rate of non-conformities in both arms. In this study, non-inferiority could not be demonstrated. Two other relevant studies are still on-going at the time of writing: START [55] and TEMPER [56]. The START study aims to investigate whether on-site visits increases the identification of major protocol deviations. All centers receive local and central monitoring, but half of the centers are randomized to receive additional annual on-site monitoring visits. In the TEMPER study, a matched-pair design is used to evaluate whether a 'targeted monitoring' approach, in which centers are continuously monitored centrally on certain performance indicators, is sensible, i.e. whether centers that are to receive an on-site monitoring visit based on a pre-specified trigger indeed have higher rates of critical issues than untriggered sites.

Note that these studies are all performed by academic parties. Although industry parties are implementing and investigating the impact of a risk-based monitoring process as well, the documentation concerning the impact assessment of those strategies is, as of yet, not sufficiently detailed. As a consequence, it is difficult to meaningfully judge the methodological validity of those assessments (e.g. [57]). We believe that there is much to gain in this respect: trials play a pivotal role in drug development, and it must be ensured that high quality is maintained and resources are adequately spend. To date, the available empirical evidence favoring one monitoring procedure over another remains limited. We therefore

suggest that more parties follow the example set by the studies outlined above. That is, design prospective studies within running trials and publish the findings and methodological details in peer-reviewed journals. From a patient, quality and cost perspective, efficient monitoring of clinical trials deserves a stronger evidence base than currently available.

5. Conclusion

Clinical trials are becoming more complex and expensive. Hence, it is necessary to critically reflect on how resources are being spent and whether the design and execution of trials can be made more efficient. One element of trial conduct that is commonly cited as a source of inefficiency is site monitoring, which traditionally relies heavily on 100% SDV and frequent on-site monitoring visits. Many authors, including regulatory agencies, have proposed alternative site monitoring strategies. Often, however, these alternatives still focus to a large extent on SDV and may benefit from a reassessment of the relation between the identified 'risks' and the proposed mitigation strategies. In addition, the number of empirical studies comparing various monitoring strategies remains limited. We believe that more effort should be put into the comparison of different monitoring procedures. Although retrospective analysis or simulation studies may provide useful insights, such comparisons are ultimately preferably performed in a prospective manner. In addition, we believe that there is room for improvement in terms of the dissemination of the findings and methodological details of such assessments.

References

1. Kleppinger CF, Ball LK. Building quality in clinical trials with use of a quality systems approach. Clinical Infectious Diseases 2010; 51(S1):S111-S116. DOI: 10.1086/653058.
2. Bhatt A. Quality of clinical trials: A moving target. Perspectives in Clinical Research 2011; 2(4):124-128.
3. FDA. Guidance for clinical trial sponsors – Establishment and operation of clinical trial data monitoring committees. 2006. URL: https://www.fda.gov/RegulatoryInformation/Guidances/ucm122046.htm
4. FDA. Guidance for industry: oversight of clinical investigations - A risk-based approach to monitoring. 2013. URL: https://www.fda.gov/RegulatoryInformation/Guidances/ucm122046.htm.
5. Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? Nature Reviews Drug Discovery 2017, 16:381-382. DOI: 10.1038/nrd.2017.70.
6. Rosenblatt M. The Changing Face of Clinical Trials - The Large Pharmaceutical Company Perspective. New England Journal of Medicine 2017; 376:52-60. DOI: 10.1056/NEJMra1510069.
7. ICH. Guideline for good clinical practice E6(R2), step 5. 2016. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002874.pdf.
8. Hines S. Targeting Source Document Verification - Targeted SDV provides data validation by comparing trial data with primary health records. Applied Clinical Trials 2011, Feb. URL: http://www.appliedclinicaltrialsonline.com/targeting-source-document-verification.
9. Institute of Medicine. Assuring data quality and validity in clinical trials for regulatory decision making - Workshop report. Washington, DC: The National Academies Press. 1999. DOI: 10.17226/9623.
10. Eisenstein EL, Lemons PW II, Tardiff BE, Schulman KA, King Jolly M, Califf RM. Reducing the costs of phase III cardiovascular clinical trials. American Heart Journal 2005; 149(3):482-488.

11. Califf RM. Clinical trials bureaucracy: unintended consequences of well-intentioned policy. Clinical Trials 2006; 3:496-502. DOI: 10.1177/1740774506073173.
12. Duley L, Antman K, Arena J, Avezum A, Blumenthal M, Bosch J, Chrolavicius S, Li T, Ounpuu S, Perez AC, Sleight P, Svard R, Temple R, Tsouderous Y, Yunis C, Yusuf S. Specific barriers to the conduct of randomized trials. Clinical Trials 2008; 5:40-48. DOI: 10.1177/1740774507087704.
13. Funning S, Grahnén A, Eriksson K, Kettis-Linblad Å. Quality assurance within the scope of Good Clinical Practice (GCP) - What is the cost of GCP-related activities? A survey within the Swedish Association of the Pharmaceutical Industry (LIF)'s members. Quality Assurance Journal 2009; 12:3-7. DOI: 10.1002/qaj.433.
14. Buyse M. Centralized statistical monitoring as a way to improve the quality of clinical data. Applied Clinical Trials 2014. Mar. URL: http://www.appliedclinicaltrialsonline.com/centralized-statistical-monitoring-way-improve-quality-clinical-data.
15. Liénard JL, Quinaux E, Fabre-Guillevin E, Piedbois P, Jouhaud A, Decoster G, Buyse M, on behalf of the European Association for Research in Oncology (AERO). Impact of on-site initiation visits on patient recruitment and data quality in a randomized trial of adjuvant chemotherapy for breast cancer. Clinical Trials 2006; 3:486-492. DOI: 10.1177/1740774506070807.
16. Bakobaki J, Joffe N, Burdett S, Tierney J, Meredith S, Stenning S. A systematic search for reports of site monitoring technique comparisons in clinical trials. Clinical Trials 2012; 9:777-780.
17. Tudor Smith C, Stocken DD, Dunn J, Cox T, Ghaneh P, Cunningham D, Neoptolemos JP. The value of source data verification in a cancer clinical trial. PLOS ONE 2012; 7(12),e51623. DOI: 10.1371/journal.pone.0051623.
18. EMA. Reflection paper on risk based quality management in clinical trials. 2013. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/11/WC500155491.pdf.

19. Grimes DA, Hubacher D, A Nanda K, Schulz KF, Moher D, Altman DG. The Good Clinical Practice guideline: A bronze standard for clinical research. Lancet 2005; 366:172-174.
20. Lörstad MH. From "too much, too late" to "right first time": Quality guru Deming's advice for clinical trials. Drug Information Journal 2000; 34:1319-1328.
21. Getz K. Low hanging fruit in the fight against inefficiency - Direction from regulatory agencies would help eradicate wasteful 100 percent source data verification. Applied Clinical Trials 2011, Mar. URL: http://www.appliedclinicaltrialsonline.com/low-hanging-fruit-fight-against-inefficiency.
22. Korieth K. The high cost and questionable impact of 100% SDV: Sponsors, CROs reluctant to alter standard practice. The CenterWatch Monthly 2011; 18(2).
23. Lörstad M. Data quality of the clinical trial process - Costly regulatory compliance at the expense of scientific proficiency. Quality Assurance Journal 2004; 8:177-182. DOI: 10.1002/qaj.288.
24. Sheetz N, Wilson B, Benedict J, Huffman E, Lawton A, Travers M, Nadolny P, Young S, Given K, Florin L. Evaluating source data verification as a quality control measure in clinical trials. Therapeutic Innovation and Regulatory Science 2014; 48(6):671-680. DOI: 10.1177/2168479014554400.
25. Olsen R, Bihlet AR, Kalakou F, Andersen JR. The impact of clinical trial monitoring approaches on data integrity and cost - a review of current literature. European Journal of Clinical Pharmacology 2016; 72:399-412. DOI: 10.1007/s00228-015-2004-y.
26. Radovich C, Frick J. Remote Source Document Verification (rSDV) - A sponsor perspective and results of implementation. Monitor 2009. URL: http://www.pharmavigilant.com/uploadDocs/1/December-Monitor-article-on-rSDV.pdf
27. Journot V, Pérusat-Villetorte S, Bouyssou C, Couffin-Cadiergues S, Tall A, Chêne G, the Optimon Collaborative Group. Remote preenrollment checking of consent forms to reduce nonconformity. Clinical Trials 2013; 10:449-459. DOI: 10.1177/1740774513480003.
28. Mealer M, Kittelson J, Thompson BT, Wheeler AP, Magee JC, Sokol RJ, Moss M, Kahn MG. Remote source document verification in two national clinical trials

networks: A pilot study. PLoS ONE 2013; 8(12):e81890. DOI: 10.1371/journal.pone.0081890.

29. Uren SC, Kirkman MB, Dalton BS, Zalcberg JR. Reducing clinical trial monitoring resource allocation and costs through remote access to electronic medical records. Journal of Oncology Practice 2013; 9:e13-e16. DOI: 10.1200/JOP.2012.000666.
30. Khosla R, Verma DD, Kapur A, Khosla S. Efficient source data verification. Indian Journal of Pharmacology 2000; 32:180-186.
31. Grieve AP. Source data verification by statistical sampling: Issues in implementation. Drug Information Journal 2012; 46:368-377. DOI: 10.1177/0092861512442057.
32. Van den Bor RM, Oosterman BJ, Oostendorp MB, Grobbee DE, Roes KCB. Efficient source data verification using statistical acceptance sampling: A simulation study. Therapeutic Innovation and Regulatory Science 2016; 50(1):82-90. DOI: 0.1177/2168479015602042.
33. Hurley C, Shiely F, Power J, Clarke M, Eustace JA, Flanagan E, Kearney PM. Risk based monitoring (RBM) tools for clinical trials: A systematic review. Contemporary Clinical Trials 2016; 51:15-27. DOI: 10.1016/j.cct.2016.09.003.
34. ADAMON risk analysis in clinical trials regarding the required amount of on-site monitoring. English version. 2009. URL: http://www.adamon.de/ADAMON_EN/Downloads.aspx.
35. TransCelerate Biopharma Inc. Risk Assessment Categorization Tool (RACT). 2014. URL: http://www.transceleratebiopharmainc.com/assets/rbm-assets/.
36. Von Niederhäusern B, Orleth A, Schädelin S, Rawi N, Velkopolszky M, Becherer C, Benkert P, Satalkar P, Briel M, Pauli-Magnus C. Generating evidence on a risk-based monitoring approach in the academic setting - lessons learned. BMC Medical Research Methodology 2017; 17, 26. DOI: 10.1186/s12874-017-0308-6.
37. TransCelerate Biopharma Inc. Position Paper: Risk-Based Monitoring Methodology. 2013. URL: http://www.transceleratebiopharmainc.com/assets/rbm-assets/
38. Djali S, Janssens S, van Yper S, van Parijs J. How a data-driven quality management systam can manage compliance risk in clinical trials. Drug Information Journal 2010; 44:359-373.

39. O'Kelly M. Using statistical techniques to detect fraud: A test case. Pharmaceutical statistics 2004; 3:237-246. DOI: 10.1002/pst.137.
40. Wu X, Carlsson M. Detecting data fabrication in clinical trials from cluster analysis perspective. Pharmaceutical statistics 2011; 10:257-264. DOI: 10.1002/pst.462.
41. Venet D, Doffagne E, Burzykowski T, et al. A statistical approach to central monitoring of data quality in clinical trials. Clinical Trials 2012: 9(6):705-713. DOI: 10.1177/1740774512447898.
42. Kirkwood AA, Cox T, Hackshaw A. Application of methods for central statistical monitoring in clinical trials. Clinicial Trials 2013; 10(5):783-806. DOI: 10.1177/1740774513494504.
43. Pogue JM, Devereaux PJ, Thorlund K, et al. Central statistical monitoring: Detecting fraud in clinical trials. Clinical Trials 2013; 10(2):225-235. DOI: 10.1177/1740774512469312.
44. Desmet L, Venet D, Doffagne E, et al. Linear mixed-effects models for central statistical monitoring of multicenter clinical trials. Statistics in Medicine 2014; 33(30):5265-5279. DOI: 10.1002/sim.6294.
45. Lindblad AS, Manukyan Z, Purohit-Sheth T, Gensler G, Okwesili P, Meeker-O'Connell, Ball L, Marler JR. Central site monitoring: Results from a test of accuracy in identifying trials and sites failing Food and Drug Administration inspection. Clinical Trials 2014; 11:205-217. DOI: 10.1177/1740774513508028.
46. Knepper D, Lindblad AS, Sharma G, et al. Statistical monitoring in clinical trials: Best practices for detecting data anomalies suggestive of fabrication or misconduct. Therapeutic Innovation and Regulatory Science 2016; 50(2):144-154. DOI: 10.1177/2168479016630576
47. Oba K. Statistical challenges for central monitoring in clinical trials: A review. International Journal of Clinical Oncology 2016; 21:28-37. DOI: 10.1007/s10147-015-0914-4
48. Timmermans C, Venet D, Burzykowski T. Data-driven risk identification in phase III clinical trials using central statistical monitoring. International Journal of Clinical Oncology 2016; 21(1):38-45. DOI: 10.1007/s10147-015-0877-5.

49. Van den Bor RM, Vaessen PWJ, Oosterman BJ, Zuithoff NPA, Grobbee DE, Roes KCB. A computationally simply central monitoring procedure, effectively applied to empirical trial data with known fraud. Journal of Clinical Epidemiology, 2017; 87:59-69. DOI: 10.1016/j.jclinepi.2017.03.018.
50. ECRIN Risk-based monitoring toolbox. URL: http://www.ecrin.org/tools/risk-based-monitoring-toolbox
51. Ansmann EB, Heckt A, Henn DK, Leptien S, Stelzer HG. The future of monitoring in clinical research - a holistic approach: Linking risk-based monitoring with quality management principles. German Medical Science 2013; 11:Doc04. DOI: 10.3205/000172.
52. Brosteanu O, Schwarz G, Houben P, Paulus U, Strenge-Hesse A, Zettelmeyer U, Schneider A, Hasenclever D. Risk-adapted monitoring is not inferior to extensive on-site monitoring: Results of the ADAMON cluster-randomised study. Clinical Trials 2017; Epub ahead of print. DOI: 10.1177/1740774517724165.
53. Journot V, Pignon JP, Gaultier C, Daurat V, Bouxin-Métro A, Giraudeau B, Preux PM, Tréluyer JM, Chevret S, Plättner V, Thalamas C, Clisant S, Ravaud P, Chêne G; Optimon Collaborative Group. Validation of a risk-assessment scale and a risk-adapted monitoring plan for academic clinical research studies - the Pre-Optimon study. Contemporary Clinical Trials 2011; 32(1):16-24. DOI: 10.1016/j.cct.2010.10.001.
54. Journot V. OPTIMON: First results of the French trial on optimisation of monitoring. Presented at Conférence Francophone d'Epidémiologie Clinique, Montpellier, 2015. URL: https://ssl2.isped.u-bordeaux2.fr/OPTIMON/docs/Communications/2015-Montpellier/OPTIMON%20-%20EpiClin%20Montpellier%202015-05-20%20EN.pdf
55. Hullsiek KH, Kagan JM, Engen N, Grarup J, Hudson F, Denning ET, Carey C, Courtney-Rodgers D, Finley EB, Jansson PO, Pearson MT, Peavy DE, Belloso WH, for the INSIGHT START Monitoring Substudy Group. Investigating the Efficacy of Clinical Trial Monitoring Strategies: Design and Implementation of the Cluster Randomized START Monitoring Substudy. Therapeutic Innovation and Regulatory Science 2015; 49(2):225-233.

56. Stenning S, Joffe N, Batra P, Jones C, Montana CD, Tappenden N, Meredith S. Update on the temper study: targeted monitoring, prospective evaluation and refinement. Trials 2013; 14(Suppl 1):P138. DOI: 10.1186/1745-6215-14-S1-P138.
57. TransCelerate Biopharma Inc. Risk-based monitoring update – volume IV. URL: http://www.transceleratebiopharmainc.com/assets/rbm-assets/.

Summary

Clinical trials serve a key role in drug development programs, by providing scientific knowledge on the risks and benefits of medical treatments or therapies. However, conducting a clinical trial is a time-consuming endeavor which typically requires a large financial budget. It is therefore important to critically reflect on how resources are being spent. Indeed, the efficiency of clinical trial conduct is becoming an increasingly active topic of research and discussion.

Estimates suggests that, typically, a considerable portion (15 to 30 percent) of a phase III trial budget is spent on the monitoring of the local investigators (i.e. the treating physicians) and their teams. Traditionally, this is the responsibility of a Clinical Research Associate (CRA), who is employed by the sponsor or contract research organization. The CRA visits the local centers on regular intervals and spends a large portion of his/her time comparing all submitted data against the source data, correcting any discrepancy that is detected.

This thesis examines and further develops alternative monitoring strategies which are aimed to conduct clinical trials more efficiently, while at the same time preserving or even improving data quality. Mainly, it focuses on methods that reduce the reliance on the physical presence of the CRA on the local centers and make better use of centrally available data, and it reflects on specific aspects of monitoring for which the effectiveness has been the subject of discussion.

The 'traditional' approach of trial monitoring, in which all submitted data is verified against source data, is known as '100% Source Data Verification (SDV)'. In **chapter 2**, a 'random sampling SDV' procedure is proposed as a possible alternative for 100% SDV. In random sampling SDV, a random selection of data points is verified against source data. The chapter describes how appropriate sample sizes can be determined and how many errors should be considered acceptable. To assess the impact of the proposed procedure in terms of the reduction in workload and the non-detected errors, a simulation study is performed. The results suggest that major reductions in workload can be achieved, while maintaining acceptable data quality levels. However, further improvement is possible, e.g. because the proposed procedure is too conservative.

Chapter 3 focuses on Central Statistical Monitoring (CSM), a form of monitoring in which digitally available data is used to detect deviating patterns. CSM can be used in parallel to on-site monitoring, and may be particularly useful for the detection of data fabrication. Data fabrication in clinical trials is, most likely, a rare phenomenon. Nevertheless, it may have serious consequences not only for the trial in which it takes place but also for the public perception towards clinical research in general. It is therefore important to actively monitor for possible data fabrication and to detect potential cases of misconduct as early as possible. This chapter presents a collection of computationally simple analyses that can be used for this purpose. When applied to empirical data from a trial in which one investigational site was known to have committed fraud, the procedure detects the corresponding center early and consistently. Application to more, independent, empirical datasets is required to assess the procedure's performance more formally.

Failure to enroll the required number of patients in a clinical trial is a common problem. In **chapter 4**, it is investigated whether it is feasible to draw an a priori distinction between centers that will meet their enrollment requirement and those that will not, based on general characteristics of the center (e.g. the type of center and the level of experience of the trial staff). Using empirical data from a large international phase III trial, the predictive value of a large set of candidate predictors on enrollment performance is assessed. The results show that the predictive value is marginal, and too limited for use in practice. Despite being based on a single trial only, these findings illustrate that care should be taken when making operational decisions based on general center characteristics. Moreover, effective site selection may not require the collection of a large number of variables, as is currently common practice.

Definitions of clinical trial outcome events may, to some extent, be subjective and therefore prone to misclassification. Adjudication committees, groups of independent experts, are employed to detect and correct misclassified events. However, various studies suggest that adjudication may have a limited impact on the substantial conclusions of clinical trials. In **chapter 5**, the impact of misclassification of outcome events on common estimators and statistical power is demonstrated by means of a number of illustrative simulations. The simulation results show that substantial problems can arise from event misclassification (even

in the context of trials where misclassification can be deemed ‘non-differential’). The decision to include an adjudication committee or not should not be based solely on the outcomes of empirical comparisons between pre- and post-adjudication results, but warrant more detailed consideration.

Chapter 6 focuses on monitoring in the specific context of pragmatic clinical trials. In pragmatic trials, more efficient alternative site monitoring procedures will be particularly useful as these trials can be expected to require even larger patient numbers and their greater reliance on centrally available data. The chapter discusses how proposed methodologies can be used in this specific context, and specific considerations that need to be addressed.

Chapter 7 provides a detailed discussion on the topic of clinical trial monitoring, the criticism on traditional approaches towards site monitoring, and the proposed alternative procedures. In addition, the available empirical evidence that compares different alternatives is discussed, as well as some general directions for future work.

Nederlandse samenvatting

Klinische trials verschaffen inzicht in de effectiviteit en mogelijke bijwerkingen van medicijnen of therapieën, en vervullen daarmee een belangrijke rol binnen de ontwikkeling van geneesmiddelen. Het uitvoeren van een trial is tijdsintensief en vergt grote financiële budgetten. Het is daarom van belang om de verschillende uitgaven binnen dit type onderzoek kritisch te evalueren. De efficiëntie waarmee medisch experimenteel onderzoek wordt verricht wordt een steeds belangrijker onderwerp van onderzoek en discussie.

Een aanzienlijk gedeelte van het totale budget van een fase III studie, naar schatting 15 tot 30 procent, wordt besteed aan het monitoren van het werk en personeel van de deelnemende ziekenhuizen. Dit monitoren is, traditioneel, primair de verantwoordelijkheid van de Clinical Research Associate (CRA), die in dienst is van de sponsor van het onderzoek of van een zogeheten 'contract research organization'. De CRA bezoekt de ziekenhuizen en besteedt een groot gedeelte van zijn/haar tijd aan het controleren van de patiëntdata zoals die aan de sponsor is doorgegeven ten opzichte van de brongegevens, waarbij eventuele discrepanties worden gecorrigeerd.

In dit proefschrift worden alternatieve monitoringstrategieën voorgesteld en onderzocht, die efficiënter zijn zonder de kwaliteit nadelig te beïnvloeden of deze zelfs te verbeteren. De primaire focus ligt op strategieën die de afhankelijkheid van de fysieke aanwezigheid van de CRA op de ziekenhuizen verminderen en waarbij meer gebruik gemaakt wordt van data die centraal beschikbaar is. Daarnaast wordt ingegaan op specifieke aspecten van het monitoren waarvan de effectiviteit onderwerp van discussie is.

De 'traditionele' vorm van monitoren, waarbij alle data door de CRA wordt vergeleken met de brongegevens, staat ook wel bekend als '100% Source Data Verification'. In **hoofdstuk 2** wordt een alternatieve procedure beschreven die uitgaat van steekproefsgewijze controle. Er wordt een methode voorgesteld waarmee de steekproefgrootte kan worden bepaald, rekening houdend met het aantal fouten in de data dat nog als acceptabel beschouwd mag worden. Om de consequenties (in termen van bespaard werk en het aantal niet-gedetecteerde fouten) van deze methode te onderzoeken is een simulatiestudie uitgevoerd. De resultaten suggereren dat er inderdaad grote besparingen mogelijk zijn zonder de kwaliteit van de data substantieel te

beïnvloeden. Verdere verbetering van de methodiek is nog mogelijk: de voorgestelde methode is bijvoorbeeld nog steeds te conservatief.

Hoofdstuk 3 is gericht op 'Central Statistical Monitoring (CSM)', een vorm van monitoren waarbij in de centraal beschikbare data wordt gezocht naar afwijkende patronen. CSM kan gebruikt worden naast een meer traditionele monitoringstrategie, en is mogelijk een effectief middel om datafabricatie te detecteren. Datafabricatie is naar alle waarschijnlijkheid een zeldzaam fenomeen. Desalniettemin kan het aanzienlijke negatieve gevolgen hebben, zowel voor het onderzoek waarin de fraude plaatsvindt als voor het publieke imago van medisch-wetenschappelijk onderzoek in het algemeen. Het is daarom van belang om actief te controleren of data mogelijk gefingeerd is, zodat potentiele fraudeurs in een zo vroeg mogelijk stadium kunnen worden gedetecteerd. In dit hoofdstuk wordt een reeks analyses beschreven dat voor dit doel kan worden gebruikt. Wanneer deze analyses worden toegepast op empirische data van een studie waarbij in één centrum daadwerkelijk patiëntgegevens zijn gefabriceerd, blijkt dat het betreffende centrum inderdaad consistent wordt gedetecteerd. Een meer formele evaluatie van de effectiviteit van de methode behoeft toepassing op meerdere onafhankelijke studies.

Een veelvoorkomend probleem in de uitvoering van medisch onderzoek is dat benodigde patiëntaantallen niet behaald worden. In **hoofdstuk 4** wordt onderzocht of het mogelijk is om, a priori, een onderscheid te maken tussen centra die wel de benodigde aantallen weten te includeren en centra waarvoor dat niet geldt, aan de hand van algemene eigenschappen van de centra (bijvoorbeeld het type centrum of de ervaring van het personeel met het uitvoeren van klinische studies). Gebruikmakend van empirische data van een internationale fase III studie wordt de voorspellende waarde van een groot aantal kandidaat-voorspellers onderzocht. De bevindingen laten zien dat deze voorspellende waarde te beperkt is voor toepassing in de praktijk. Hoewel slechts gebaseerd op data van één studie, laten deze bevindingen zien dat er voorzichtig moet worden gehandeld wanneer er operationele beslissingen worden genomen op basis van algemene eigenschappen van deelnemende centra. Het illustreert daarnaast mogelijk ook dat de huidige praktijk, waarin een groot aantal factoren wordt meegewogen in

de beslissing om een centrum wel of niet te includeren in een studie, wellicht maar in beperkte mate effectief is.

De definities die worden gehanteerd om een klinische uitkomst te classificeren kunnen, tot op zekere hoogte, subjectief van aard zijn. Daarmee is het mogelijk dat bepaalde typen uitkomsten verkeerd gediagnosticeerd worden. Tijdens het uitvoeren van een klinische studie kan een onafhankelijke groep experts (een zogeheten 'adjudication committee') in het leven worden geroepen om mogelijke verkeerde diagnosen te detecteren en te corrigeren. De resultaten van meerdere onderzoeken lijken echter te suggereren dat dit proces slechts van beperkte invloed is op de resultaten van een klinische trial. In **hoofdstuk 5** wordt, aan de hand van illustratieve simulaties, besproken wat het effect van misdiagnose kan zijn op veelgebruikte schatters en op de statistische power van significantietoetsen. Er wordt geconcludeerd dat misdiagnose wel degelijk een substantieel effect kan hebben, ook in een klinische trial waar de kans op misdiagnose gelijk is in de twee onderzoeksarmen. De keuze om wel of geen zogeheten adjudication committee aan te stellen vereist daarom een gedegen discussie, en dient niet enkel gebaseerd te worden op de uitkomsten van empirische vergelijkingen.

Hoofdstuk 6 richt zich op monitoren in de specifieke context van pragmatische klinische trials. Vanwege de mogelijk nog hogere patiëntaantallen en de afhankelijkheid van centraal aanwezige patiëntgegevens is de noodzaak tot meer efficiënte vormen van monitoren, in dit type trial in het bijzonder, groot. In dit hoofdstuk wordt besproken op welke wijze voorgestelde alternatieven kunnen worden toegepast in deze context.

Hoofdstuk 7 is een uitgebreide discussie over het monitoren van klinische trials, de kritiek op de 'traditionele' vormen van monitoren, en alternatieve monitorprocedures zoals die zijn voorgesteld als gevolg van die kritiek. Daarnaast bespreekt het de mate waarin alternatieve strategieën voor het monitoren van klinische trials empirisch getoetst of vergeleken zijn, en bevat het algemene suggesties voor vervolgonderzoek.

Dankwoord

Prof. dr. C.B. Roes, geachte promotor, beste Kit, onze discussies waren voor mij altijd motiverend. Ik heb veel bewondering voor je kennis en ervaring met betrekking tot medisch onderzoek, en je vindingrijkheid voor wat betreft het bedenken van oplossingen voor complexe methodologische problemen. Ik heb veel geleerd van jouw inzichten en ideeën en hoop dat ik dat in de toekomst kan blijven doen. Bedankt voor het geduld, vertrouwen, de prettige samenwerking, en dat ik altijd bij je terecht kon met mijn vragen.

Prof. dr. D.E. Grobbee, geachte promotor, beste Rick, bedankt voor het vertrouwen en het geven van de mogelijkheid om, naast het werk aan deze thesis, ook onderdeel te zijn van Julius Clinical. Werken bij Julius Clinical heeft mij de mogelijkheid gegeven om kennis op te doen met betrekking tot de complexiteiten omtrent de uitvoering van trials. Jouw ervaring, suggesties voor onderzoeksonderwerpen en ideeën over de invulling daarvan hebben me enorm geholpen in het vervaardigen van deze thesis. Hartelijk dank daarvoor.

Dr. B.J. Oosterman, geachte co-promotor, beste Bas, bedankt voor de prettige samenwerking. Jouw kennis en je kritische, gestructureerde blik hebben de artikelen in dit proefschrift veel goeds gedaan: inhoudelijk, maar ook in termen van leesbaarheid. Zelfs na je afscheid bij Julius Clinical kon ik altijd rekenen op advies als ik dat vroeg. Heel erg bedankt daarvoor!

Many co-authors have contributed to the research presented in this thesis. I thank all of the authors for the discussions and commitment to our projects. Martijn: Bedankt voor je hulp bij de data-analyse in hoofdstuk 2. Peter, bedankt voor de nuttige discussies met betrekking tot hoofdstuk 3. Tim: Ik heb veel geleerd van onze besprekingen aangaande hoofdstuk 3, 4, en 5. Paco: Bedankt voor de prettige samenwerking in het GetReal project (hoofdstuk 6).

To the members of the reading committee: prof. dr. H. Boersma, prof. dr. M.J.C. Eijkemans, prof. dr. I.G. Klugkist, prof. dr. M. Landray, and prof. dr. K.G.M. Moons. I thank you for your time and effort in reading and assessing this thesis.

To current and former colleagues of the department of Biostatistics and Research Support: Bert, Caroline, Cas, Esther, Hae-Won, Ingeborg, Kit, Julien, Konstantinos, Maarten, Marijn,

Paul, Peter, Putri, Rebecca, René, Rik, Stavros, Victor, and Willem, and to other colleagues at the Julius Center: Katrien, Marian, Romin, Shona, Thomas, Timo, Valentijn: It is a great pleasure working with you all. René: Bedankt voor het geven van de mogelijkheid om mijn werk voort te kunnen zetten binnen de afdeling Biostatistiek. Willem: bedankt voor de begeleiding tijdens mijn masterscriptie. Ik heb veel van je geleerd.

Ook wil ik alle (oud-)collega's bij Julius Clinical bedanken. In het bijzonder mijn kamergenoten en natuurlijk de collega's die direct of indirect betrokken zijn geweest bij het onderzoek in deze thesis: Aize, Anna, Bas (x2), Elly, Ferdi, Giene, Hans, Helma, Loes, Lotte, Maaike, Marijke, Martijn, Nastaran, Nel, Patricia, Peter, Richard, Rick, Ronald, Saskia, Suzanne, Tim, Wai Ming, Wanda en Wendela.

I wish to thank Boehringer Ingelheim for providing the opportunity to use the ESPS2 trial data in chapter 3.

Lieve familie en vrienden, bedankt, zonder jullie had deze thesis nooit tot stand kunnen komen.

Paranimfen, Guus en Rolf, ik ben trots dat jullie op deze dag naast me willen staan.

Schoonfamilie, Jop, Wilma, Silke, Hasse, Tim en Jurre, bedankt dat ik me bij jullie altijd welkom kan voelen, en voor jullie interesse en betrokkenheid.

Margriet, ma, ik ben jou en pa enorm dankbaar voor jullie steun en interesse. Hilde, Henk, Emiel: datzelfde geldt natuurlijk voor jullie. De afgelopen jaren waren moeilijk, maar ik ben heel erg trots op hoe sterk jullie zijn geweest.

Lieve Diede, onderschat niet hoeveel je hebt bijgedragen aan dit werk: zonder jou was dit niet gelukt. Heel erg bedankt voor je luisterend oor en voor je hulp, interesse, humor en zorgzaamheid.

Curriculum Vitae

Rutger van den Bor was born on August 30, 1988 in Ermelo, the Netherlands. As of 2007, he studied Sociology at the University of Utrecht. Following his graduation in 2010, he started the two-year master program Methodology and Statistics of Behavioral and Social Sciences at the University of Utrecht, graduating as Master of Science in 2012. As part of this program, he wrote his master's thesis during an internship at the department of Biostatistics and Research Support of the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht. The thesis investigated the performance of two procedures (inverse probability weighting and G-computation) that can be used to adjust for confounding in marginal causal effect modelling. In December 2012, Rutger started his PhD project at Julius Clinical Research Ltd. and the Julius Center for Health Sciences and Primary Care, under supervision of prof. dr. C.B. Roes, prof. dr. D.E. Grobbee and dr. B.J. Oosterman. The research resulting from this project is presented in this thesis. In 2015, he graduated as Master of Science in Epidemiology with a specialization in Medical Statistics at the University of Utrecht. As of July 2016, he works as a statistician and researcher at Julius Clinical Research Ltd. and the department of Biostatistics and Research Support of the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht.