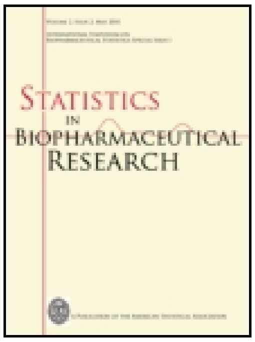
Use of the Beta-Binomial Model for Central Statistical Monitoring of Multicenter Clinical Trials

Peer-reviewed author version

Desmet, Lieven; Venet, David; Doffagne, Erik; Timmermans, Catherine; LEGRAND, Catherine; BURZYKOWSKI, Tomasz & BUYSE, Marc (2017) Use of the Beta-Binomial Model for Central Statistical Monitoring of Multicenter Clinical Trials. In: STATISTICS IN BIOPHARMACEUTICAL RESEARCH, 9(1), p. 1-11.

DOI: 10.1080/19466315.2016.1164751
Handle: http://hdl.handle.net/1942/24156



Taylor & Francis
Taylor & Francis Group

Statistics in Biopharmaceutical Research

ISSN: (Print) 1946-6315 (Online) Journal homepage: http://www.tandfonline.com/loi/usbr20

Use of the beta-binomial model for central statistical monitoring of multicenter clinical trials

Lieven Desmet, David Venet, Erik Doffagne, Catherine Timmermans, Catherine Legrand, Tomasz Burzykowski & Marc Buyse

To cite this article: Lieven Desmet, David Venet, Erik Doffagne, Catherine Timmermans, Catherine Legrand, Tomasz Burzykowski & Marc Buyse (2016): Use of the beta-binomial model for central statistical monitoring of multicenter clinical trials, Statistics in Biopharmaceutical Research, DOI: 10.1080/19466315.2016.1164751

To link to this article: http://dx.doi.org/10.1080/19466315.2016.1164751

Accepted author version posted online: 06 Apr 2016.

Use of the beta-binomial model for central statistical monitoring of multicenter clinical trials

Lieven Desmet*
ISBA, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
David Venet
IRIDIA, Université Libre de Bruxelles, Brussels, Belgium
Erik Doffagne
CluePoints S.A., Mont-Saint-Guibert, Belgium
Catherine Timmermans
ISBA, Université Catholique de Louvain, Louvain-la-Neuve &
Département de Mathématique, Université de Liège, Belgium
Catherine Legrand
ISBA, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
Tomasz Burzykowski
I-BioStat, Hasselt University & IDDI S.A., Louvain-la-Neuve, Belgium
Marc Buyse
I-BioStat, Hasselt University, Belgium & IDDI Inc., San Francisco, CA, USA

Abstract

As part of central statistical monitoring of multicenter clinical trial data, we propose a procedure based on the beta-binomial distribution for the detection of centers with atypical values for the probability of some event. The procedure makes no assumptions about the typical event proportion and uses the event counts from all centers to derive a reference model. The procedure is shown through simulations to have high sensitivity and high specificity if the contamination rate is small and the atypical event proportions are the result of some systematic shift in the underlying data generating mechanism.

Keywords: beta-binomial, central statistical monitoring, multicenter clinical trial, error detection

*The authors gratefully acknowledge financial support from the Walloon Government under the BioWin framework (Consortium agreement nr. 6741). Partial support for this research (Catherine Legrand and Tomasz Burzykowski) was provided by the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy). The authors thank two reviewers for making suggestions that significantly improved the paper.

1 Introduction

Central Statistical Monitoring (CSM) is a novel approach to the monitoring of multicenter clinical trial data, based on statistical tests, models, and scoring algorithms [1, 2, 3, 4]. The rationale behind CSM is that data quality issues arising from transcription errors, measurement problems, misunderstandings, procedural issues, data tampering or even data fabrication may remain undetected with on-site monitoring, but are more easily detected when the data of each center are compared with the data from all other centers. Some authors propose to use statistical tests to identify centers with unusual data patterns for selected clinical variables [2, 4], while others have used trials with known cases of fraud to build a model predictive of fraud based on a limited number of key clinical variables [3]. In the most radical implementation of CSM, a large battery of statistical tests is applied to all variables collected in the clinical database and the resulting *p*-values are combined in an overall *data inconsistency score* for each center [1].

In this paper, we focus on the test for the comparison of proportions in the aforementioned context. The test is one of the most frequently used in CSM, because it applies to all situations where a binary variable captures an event that is either directly measured (e.g., if the patient reported a specific adverse event) or derived from other data (e.g., if there are observations missing for a given variable). Assuming that the trial procedures are similar across the centers (as imposed by the protocol) and that patient populations are comparable, we expect similar proportions of such events in all centers. However, we aim at detecting situations where an issue in one of the centers leads to a different data-generating mechanism, resulting in a lower or higher proportion than in other centers. We will refer to such a center as *atypical*, in contrast to the *typical* centers (not affected by any issue).

Our test procedure is based on the idea of using the beta-binomial model to account for extra-binomial variation often seen in biological and biomedical data (see, e.g., Griffiths [7] and Williams [8]). Chuang-Stein [5] followed this approach to detect atypical adverse event rates in new trials based on a meta-analysis of comparable historical trials. To this aim, she used the beta-binomial model estimated from the pooled data as a reference.

In the CSM setting, we aim at automatically detecting atypical centers within a single trial without any prior information. Moreover, we do not wish to rely on any prior assumptions about the nature of the binary event or its probability. We do assume, however, that observed event counts in the typical centers are realizations of a single data-generating mechanism, namely a beta-binomial model with (unknown) location parameter μ_0 and (small) overdispersion parameter ρ_0. In addition, we assume that

(i) atypical centers, if present, represent only a small fraction of all centers (say, at most 5% and very often a single center), and

(ii) event counts in the atypical centers are realizations from a beta-binomial model with a different (unknown) location parameter μ_1 and (small) overdispersion parameter ρ_1.

In our experience, the extra-binomial variation is usually small and can be adequately captured by values typically in the 0.01 to 0.05 range for parameters ρ_0 and ρ_1.

The detection problem consists of automatically detecting the atypical centers, if any. In Section 2 we discuss the rationale for a simple detection procedure based on beta-binomial modeling and evaluate its performance in terms of *power* (the ability to detect atypical centers) and *specificity* (the ability to avoid flagging typical centers). To enhance the performance and obtain more consistent properties, a number of pre-processing and adjustment steps are introduced, leading to the comprehensive algorithm presented in Section 3. A simulation study, as well as real examples, are presented in Section 4. We end with a discussion of the properties and possible applications of the proposed approach for central statistical monitoring of multicenter clinical trials. Technical details on the beta-binomial distribution are deferred to the Appendix. All simulations were carried out using the R software.

2 Beta-binomial-modeling-based detection approach

To fix notations, we consider grouped binary data, where the raw data are pairs (x_i, n_i) or proportions x_i/n_i for $i = 1, ..., N$, with x_i denoting the event counts, n_i the corresponding number of trials and N the number of groups. For convenience, we will refer to the groups as *centers*, without assuming that n_i always corresponds to a number of subjects, allowing for applications where multiple items are observed per subject.

We assume that the raw dataset consists of N_0 *typical* centers with counts $x_i \sim \text{Beta-binomial}(n_i, \mu_0, \rho_0)$ (*the null model*), and possibly, in addition, N_1 *atypical* centers with counts $x_i \sim \text{Beta-binomial}(n_i, \mu_1, \rho_1)$ (*the alternative model*), with $\mu_1 \neq \mu_0$ and small ρ_0 and ρ_1. Without loss of generality we assume throughout the text that $\mu_0 \leq 0.5$ (the *symmetry property*).

The detection procedure assesses each center in terms of the plausibility of its observed count, with respect to the (unknown) null-distribution that is assumed valid for the typical centers.

The p-value for an observed count x in a center of size n is defined as follows [9]:

$$p(x) = \begin{cases} \min(2P(X \geq x), 1) & \text{if } x > n\mu_0 \\ \min(2P(X \leq x), 1) & \text{if } x \leq n\mu_0 \end{cases} \tag{1}$$

where, under the null model, $X \sim \text{Beta-binomial}(n, \mu_0, \rho_0)$.

We use this p-value in the following decision rule: *flag* an observation x as suspicious if and only if $p(x) < \alpha_{crit}$. Since the Type I error probability equals α_{crit}, we will refer to this value as the *significance level* (taken 0.05 throughout this text).

In practice, we have no means of deriving the null model from the observed data, because some of the centers in our dataset may be atypical. Therefore, we will use as a working reference the model estimated from *all* data, the so-called *hybrid* model, and assume that it is a good approximation of the null-model, provided that the *contamination rate*, defined as $N_1/(N_0 + N_1)$, is low.

Based on the aforementioned assumptions, a simple detection procedure is defined as follows:

1. Fit a beta-binomial model to all data $(x_i, n_i)_{i=1,\ldots,N}$.
2. Based on the estimated (hybrid) model, assign to each center i a p-value p_i and flag it if $p_i < \alpha_{crit}$.

The procedure is illustrated in a simulated example with 48 typical centers generated from $\text{Beta-binomial}(n = 50, \mu_0 = 0.3, \rho_0 = 0.02)$, and 2 atypical centers from $\text{Beta-binomial}(n = 50, \mu_1 = 0.6, \rho_1 = 0.02)$. Relevant densities and the resulting estimated hybrid model are depicted in the left panel of Figure 1. Note that, in this case the contamination rate is small (4%) and the hybrid model is a reasonable approximation of the null model.

The decision rule for detection of an atypical center of size 50, based on the estimated hybrid model, is illustrated in the right panel of Figure 1. In particular, centers with at most 5 or at least 27 events are considered inconsistent with the null model. This implies that most centers consistent with the alternative model would be flagged.

The left panel of Figure 1 presents also the estimated hybrid model for the contamination rate of 40%, i.e., for the case of 30 typical and 20 atypical centers. The hybrid-model-based critical bounds for detecting an atypical center are equal to 5 and 39. However, in that case, the atypical centers will not always be detected. This is because the hybrid model does not offer a reasonable approximation of the null model. In particular, as compared to the null model, the hybrid model has a larger variance and a smaller critical region. As a result, the specificity is conservatively controlled at $1 - \alpha_{crit}$, i.e., typical centers are mostly not flagged.

While the procedure seems promising, its performance is not guaranteed under all circumstances. In fact, in another example with a contamination rate of 4% and overdispersion $\rho_0 = \rho_1 = 0.02$, but with $\mu_0 = 0.001$ and $\mu_1 = 0.999$, the power of the detection procedure is surprisingly equal to 0, in spite of the larger difference $|\mu_0 - \mu_1|$. The reason is that the estimated hybrid density becomes U-shaped and it does not allow detecting of atypical centers, as shown in Figure 2. This situation corresponds to a violation of the assumption that the hybrid model is a reasonable approximation of the null model. To recover the power in such cases, in the next section we propose a refined procedure that includes a model-adjustment step. Moreover, we provide a comprehensive algorithm for practical application in an automated context such as CSM.

3 A practical detection procedure

The algorithm as detailed below follows the ideas of the beta-binomial-based detection approach but includes preliminary steps to deal with particular cases and the beta-binomial model fitting is made more robust in the sense that it uses method of moments estimation as a backup if convergence fails for maximum likelihood. At several points it is assessed whether a binomial model is more appropriate and if a beta-binomial model is selected, model adjustment is performed if needed.

3.1 The full algorithm

Stage 1: Preliminary steps.

- 1A. Stop the procedure if $x_i = 0$ in all centers or $x_i = n_i$ in all centers.
- 1B. Define and compute the following quantities:

$$\hat{p} := \frac{1}{N}\sum_{i=1}^{N}\frac{x_i}{n_i} \tag{2}$$

$$\hat{p}_w := \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} n_i}; \tag{3}$$

$$\hat{\mu}_M := \hat{p}; \tag{4}$$

$$\hat{\rho}_M := \frac{\sum_{i=1}^{N}(\frac{x_i}{n_i} - \hat{p})^2 - \hat{p}(1-\hat{p})[\sum_{i=1}^{N}\frac{1}{n_i}(1-\frac{1}{N})]}{\hat{p}(1-\hat{p})[\sum_{i=1}^{N}(1-\frac{1}{N}) - \sum_{i=1}^{N}\frac{1}{n_i}(1-\frac{1}{N})]} \tag{5}$$

$$S := \sum_{i=1}^{N}\frac{(x_i - n_i\hat{p}_w)^2}{\hat{p}_w(1-\hat{p}_w)}; \tag{6}$$

$$Z := \frac{S - \sum_{i=1}^{N} n_i}{\sqrt{2\sum_{i=1}^{N} n_i(n_i-1)}}. \tag{7}$$

- 1C. Go to Stage 2C if $Z < Q_{N(0,1)}(0.95)$ and $\hat{\rho}_M < 10^{-3}$, where $Q_{N(0,1)}(0.95)$ is the 95% percentile of the standard-normal distribution.

Stage 2: Model selection and estimation.

Set δ to a very small value, e.g., $\delta = 10^{-6}$. Denote the maximum-likelihood estimates of the Beta-binomial parameter by $(\hat{\mu}_L, \hat{\rho}_L)$, and the iterated method-of-moments estimates (as in reference [5]) by $(\hat{\mu}_I, \hat{\rho}_I)$.

- 2A. Attempt maximum likelihood estimation with starting values $(\hat{\mu}_M, \hat{\rho}_M)$.
  - In case of convergence, and if $\hat{\rho}_L \in [\delta, 1-\delta]$: keep $(\hat{\mu}_L, \hat{\rho}_L)$ and go to Stage 3.
  - In case of convergence, and if $\hat{\rho}_L < \delta$: if $Z < Q_{N(0,1)}(0.95)$ or $\hat{\rho}_M < 10^{-3}$ go to 2C; otherwise, go to 2B.
  - In case of convergence, and if $\hat{\rho}_L > 1-\delta$: set $\hat{\rho}_L$ to $1-\delta$.

  - In case of non-convergence: compute $(\hat{\mu}_I, \hat{\rho}_I)$ and, if $Z < Q_{N(0,1)}(0.95)$ or $\hat{\rho}_I < 10^{-3}$. go to 2C; otherwise, go to 2B.

- 2B. Keep $(\hat{\mu}_I, \hat{\rho}_I)$ and go to Stage 3.

- 2C. Use the binomial model based on $\hat{p}_w$ and go to Stage 4.

Stage 3: Model adjustment step (for the estimated beta-binomial model in terms of $\hat{\alpha}$ and $\hat{\beta}$).

- Compute, with the following, suggested default values $(\gamma = 0.1, \alpha_{target} = \beta_{target} = 1)$

$$(\tilde{\alpha}, \tilde{\beta}) = \begin{cases} (\hat{\alpha}, (1 - \lambda(\hat{p}))\hat{\beta} + \lambda(\hat{p})\beta_{target}) & \text{if } \hat{p} < 0.5 \text{ and } \hat{\beta} < 1 \\ ((1 - \lambda(\hat{p}))\hat{\alpha} + \lambda(\hat{p})\alpha_{target}, \hat{\beta}) & \text{if } \hat{p} \geq 0.5 \text{ and } \hat{\alpha} < 1 \end{cases}$$

where the function $\lambda : [0,1] \to [0,1]$ is defined as:

$$\lambda(x) = \begin{cases} (\cos(\pi\, x/\gamma) + 1)/2 & \text{if } x < \gamma \text{ or } x > 1 - \gamma \\ 0 & \text{otherwise} \end{cases}$$

- Assign the adjusted model $(\tilde{\alpha}, \tilde{\beta})$ to centers with $x_i > n_i - x_i$ for $\hat{p} < 0.5$ or centers with $x_i < n_i - x_i$ for $\hat{p} \geq 0.5$ (centers that deviate strictly from the overall tendency). Assign the unadjusted model (from Stage 2) to the remaining centers.

Stage 4: Evaluation of individual centers based on the selected model.

- Compute p-values using formula (1) based on the selected model.

- Flag center i if its p-value $p_i < \alpha$.

3.2 Further details and methodological comments

Stage 1A abandons the test when data are degenerate due to limited center sizes with μ_0 and μ_1 both close to 0 (or 1).

In Stage 1B, we compute useful characteristic quantities. The mean estimates $\hat{p}$ and $\hat{p}_w$ are special cases of Kleinman's [11] moment estimator $\hat{\mu}$ (see equation (8) below) when ρ is equal to 1, and 0, respectively. They coincide in the balanced case, but when overdispersion is present and/or centers have unequal size, $\hat{p}$ is more accurate. Kleinman's estimates $\hat{\mu}_M = \hat{p}$ and $\hat{\rho}_M$ (taking

$\rho = 1$ in equation (9) below) serve as starting values for Stage 2A. Quantities S and Z are related to Tarone's test [15] for a binomial model against a beta-binomial alternative.

Stage 1C aims at preventing a beta-binomial fit if there is strong evidence (based on the combination of Tarone's test and a method of moments estimate of the overdispersion) that the use of a binomial model is justified.

Maximum likelihood estimation (see, e.g., [6] or [10]) with starting values ($\hat{\mu}_M$,$\hat{\rho}_M$) is attempted in Stage 2A. If it fails to converge, we rely on the iterated method of moments estimates, as described in reference [5], by successively iterating equations (8) and (9):

$$w_i := \frac{n_i}{1+\rho(n_i-1)}; \quad \hat{\mu} := \frac{\sum_{i=1}^N w_i \frac{x_i}{n_i}}{\sum_{i=1}^N w_i} \tag{8}$$

$$\hat{\rho} := \frac{\sum_{i=1}^N w_i(\frac{x_i}{n_i} - \hat{\mu})^2 - \hat{\mu}(1-\hat{\mu})[\sum_{i=1}^N \frac{w_i}{n_i}(1 - \frac{w_i}{\sum_{i=1}^N w_i})]}{\hat{\mu}(1-\hat{\mu})[\sum_{i=1}^N w_i(1 - \frac{w_i}{\sum_{i=1}^N w_i}) - \sum_{i=1}^N \frac{w_i}{n_i}(1 - \frac{w_i}{\sum_{i=1}^N w_i})]} \tag{9}$$

The iterative procedure that is an extension of Kleinman's method [11] and is recommended by Chuang-Stein [5]. In our experience, the maximum likelihood approach shows a more consistent performance across all scenarios of interest and has the advantage of adapting well to unbalanced setups. We used the R software implementation provided by Yee (function `vglm` in package VGAM, [12] and [13]) and used starting values as recommended by Yee to avoid convergence problems [14].

If $\hat{\rho}_L$ estimates turn out to be extremely close to 0 or 1, additional steps are carried out to safeguard the specificity of the procedure.

In Stage 3, we address the power issue described in Section 2. In terms of the assumptions stated before, the issue occurs when (*i*) the contamination rate is small and (*ii*) μ_0 tends to 0 while μ_1 tends to 1. In such circumstances the estimated hybrid model may have a large overdispersion parameter, thus violating the assumption that it is a reasonable approximation of the null model. In this case, we take a more pragmatic position, attempting to restore the performance of the procedure by adjusting the hybrid model in such a way that detection is enabled while specificity is maintained.

In practice, we rely on $\hat{p}$ (the tendency of the majority of the centers) and the shape parameters

$\hat{\alpha}$ and $\hat{\beta}$ of the estimated hybrid model to assess the need for adjustment. In particular, when $\hat{p}$ is close to 0 (which implies that $\hat{\alpha} \ll \hat{\beta}$) and $\hat{\beta} < 1$, the adjustment is needed. An *adjusted* beta-binomial model with shape parameters $\tilde{\alpha} = \hat{\alpha}$ and $\tilde{\beta} > 1$ instead of $\hat{\beta}$ enables detection of centers with large event probabilities. On the other hand, with increasing $\hat{p}$, the mismatch between the null model and the hybrid model becomes less severe for small contamination rates. With $\hat{p}$ sufficiently far from 0 and 1, the problem disappears as the hybrid model becomes unimodal.

To achieve a smooth transition between the situations that require adjustment and those that do not require it, we propose to obtain $\tilde{\beta}$ by interpolation between the original value $\hat{\beta}$ and the target value, i.e., 1 (though larger values are possible), where the coefficient of interpolation is $\lambda(\hat{p})$, as defined before. The function λ is inspired by the split cosine-bell function used for tapering in spectral density estimation (see, e.g., Bloomfield [16]) and is depicted in Figure 3 for two values of the parameter γ. Note that λ is zero on the $[\gamma, 1-\gamma]$ interval. Thus, the adjustment has no effect when $\hat{p}$ lies away from 0 or 1 by a distance larger than γ.

The adjusted density and beta-binomial distribution for the example shown in Figure 2 are depicted in Figure 4. The adjustment enables detection of centers with at least 14 events in this example.

4 Simulation study

In this section, we assess the properties of the proposed procedure, both on simulated data and on real-life clinical trials.

4.1 Motivation and setup

As an analytical approach is intractable, we conducted extensive simulations to investigate the performance of the algorithm.

A large number of scenarios in terms of the parameters (μ_0, μ_1) were considered, because the detection problem depends not only on the difference in the location parameters, but also on their individual values. However, because of the symmetry property, it is sufficient to focus on the

$(0, 1) \times (0, 0.5]$ region, which we covered by a grid of sufficient resolution to fully capture the behavior of the procedure (especially near the boundaries).

Scenarios were then further defined by the assumed amount of overdispersion (for the sake of simplicity, we assumed $\rho = \rho_0 = \rho_1$), the contamination rate, and the sizes of typical and atypical centers.

Our simulations consisted of a large number of trials in which events were generated at random in the N_1 atypical centers according to the Beta-binomial(n_i, μ_1, ρ) distribution, and in the $N - N_1$ typical centers according to the Beta-binomial(n_i, μ_0, ρ) distribution. For each simulated trial, the algorithm of Section 3 was applied. While we were mainly interested in small contamination rates for signal detection, we also wanted to assess the properties of the algorithm, and especially its specificity, for larger contamination rates. The outcomes in individual simulations were interpreted in terms of numbers of true positive (TP), false negative (FN), false positive (FP) and true negative (TN) findings. Power and specificity were computed as $TP/(TP + FN)$ and $FP/(FP + TN)$, respectively. As the standard error on an estimate $\hat{\pi}$ of the power or specificity based on N_{sim} replications is approximately equal to $\sqrt{\hat{\pi}(1 - \hat{\pi})/(N_1 N_{sim})}$, we adjusted the number of replications N_{sim} in function of the level of the contamination rate to keep the denominator constant. As a result, for the considered parameter configurations, the standard error was equal to at most 0.011.

We will refer to the signals ($\mu_1 \neq \mu_0$) as *departures*. If $\mu_0 < 0.5$, we refer to 0 as the *adjacent* boundary and 1 as the *opposite* boundary.

4.2 Overdispersion and contamination effects

The first simulation study (see Table 1) was carried out in a balanced setup. While this setting is not necessarily representative of clinical trials, it is a good starting point to assess the effects of overdispersion and contamination in the absence of effects due to small sample size. The results will be used as a benchmark in further evaluations.

Figure 5 shows the plots of power *versus* μ_1 for $\mu_0 = 0.5$ (top row) and 0.01 (bottom row) combined with two levels of overdispersion: the limit value of 0 (binomial setup) and 0.1 (a fairly large amount of overdispersion). The power curves form an ordered bundle as a function of the contamination rate: the curve for the lowest contamination rate is on top and that for the highest contamination rate at the bottom.

For $\mu_0 = 0.5$, the curves are symmetric and we can distinguish different trends: with small contamination rates (say up to 10%), the power increases rapidly to 1 with increasing $|\mu_0 - \mu_1|$, but with 20% contamination the power decreases to 0 at the boundary; with more than 40% contamination rate the curves become flatter.

For $\mu_0 = 0.01$, only departures to the opposite boundary are detected and the power curve is not monotonic at 10% contamination.

By comparing the left and right panel in each row we can assess the effect of increasing overdispersion. As expected, the detection becomes more difficult. The power curves in the left-hand panels ($\rho = 0$, binomial model) are systematically higher than their counterparts in the right-hand side panels, all other parameters being the same.

In addition, when $\mu = 0.01$ and $\rho = 0.1$, the proposed algorithm and the beta-binomial-model-based (”unadjusted”) procedure (see Section 2) show different behavior (power for the latter shown with a dashed curve). In particular, the adjustment guarantees a monotonic power curve for departures to the opposite boundary at 5% contamination. Specificity of the procedure is conservatively controlled at the 5% level, as can be concluded from the results reported in Table 3.

Based on the simulation exercise we could conclude that the adjusted procedure shows a consistent behavior, irrespectively of (a small amount of) contamination.

4.3 Unbalanced setup and size effects

The next series of simulation studies was carried out in an unbalanced setup with only 20 centers, a more realistic setting in the context of e.g. Phase II trials. To simulate centers of different sample sizes, we considered empirical distributions of the number of patients observed in three multicenter clinical trials (Figure 6). The distributions are right skewed (small sizes are dominant) and have very different ranges. The lower and upper quartiles are equal to 6 and 27 for Distribution 1; 51 and 151 for Distribution 2; and 4 and 9 for Distribution 3.

The general setup for the simulations is given in Table 2. Center sample sizes were generated at random from the distributions shown in Figure 6. In some simulations, in order to control the sample sizes of the atypical centers, they were assumed to be fixed at a quantile of the distribution.

First, we consider simulations, in which the sample sizes for all centers were drawn at random from Distribution 2 (see Figure 6).

To visualize the performance of the proposed algorithm, we consider a grid of values for μ_0 and μ_1 ranging, respectively, from 0 to 0.5 and from 0 to 1 in steps of 0.02. Note that 0 and 1 were replaced by 10^{-6} and $1 - 10^{-6}$, respectively. Figure 7 presents heatmaps summarizing the power and specificity of the adjusted algorithm (Section 3.1), together with the power of the beta-binomial-model based procedure (Section 2). All heatmaps in this Section relate to the settings of $\rho = 0.01$ and 5% contamination. With a value $\rho = 0.1$ which is deemed to be a fairly large overdispersion (see Figure 12), power is mildly affected (as in Figure 5) but the specificity stays high (see Table 3)

The power of the adjusted algorithm is depicted in the left-hand-side panel of Figure 7. Along the diagonal $\mu_0 = \mu_1$, the false-positive detection rate is at most 5% and the power is monotonically increasing for increasing $|\mu_1 - \mu_0|$, as expected. The specificity, depicted in the right-hand-side panel, is controlled at 95%.

The middle panel shows the power obtained for the same simulated data for the beta-binomial-model procedure: a difference is seen in the scenarios where $\mu_0 = 10^{-6}$ and μ_1 approaches the opposite boundary.

In the next series of simulations, sample sizes for the typical centers were random, but sample sizes of the atypical centers were fixed at the median or the 5%-tile of the sample size distributions from Figure 6. The heatmaps showing the resulting power of the adjusted algorithm are shown in Figures 8 and 9, respectively. Note that, as compared to Figure 7, the graphs have a coarser resolution in the interior (steps of 0.05).

The adjusted procedure was performing best for Distribution 2 and worst for Distribution 3, where the power attained for μ_1 at the boundary was smaller (especially with μ_0 close to 0.5) and there was a very limited power to detect departures to the adjacent boundary.

When atypical-center sample sizes were small, the power was high for a large overall center-specific sample size (Distribution 2, middle panel). But for the other two distributions, for which the atypical-center sample size was only 2, adequate power was maintained only when μ_0 was rather small (say up to 0.1) and was reduced to 0 with more central values of μ_0. The detection of departures to the opposite boundary for $\mu_0 = 10^{-6}$ was again enabled by the adjustment step, as can be seen by comparing Figure 9 and Figure 10.

The specificity across all scenarios and all simulations is summarized in Table 3. Overall, the specificity was adequately controlled. Only in few scenarios the specificity fell below 95%; these were the more challenging settings where the contamination rate was high (40% and 50%) and $\rho = 0$.

4.4 Actual clinical trials

We illustrate the use of the proposed detection procedure using data from two clinical trials.

The first trial was carried out in 37 centers. We focus on the missingness of a laboratory value that had to be obtained for each patient at a number of visits over time. The proportion of missing lab-values per center is shown in the left-hand-side panel of Figure 11. There was a considerable amount of missing data at all centers because of a delay in entering the laboratory values in the database: the overall proportion of missing values was around 11% ($\hat{p} = 0.11253$). In one of the centers 165 out of 866 (19.05%) records were missing, and this center (indicated by the cross the left-hand-side panel of Figure 11) was flagged with a p-value of 0.00227. In this trial, the overdispersion was very small ($\hat{\rho} = 0.0038$) and the hybrid distribution reflected the null distribution quite well (under the assumption of a small contamination rate). However, the beta-binomial model appears preferable to the binomial model, which would have flagged the outlying center with a p-value of 1.47×10^{-11}, which is quite extreme, and in addition the binomial model is likely to generate an excessive number of false positive findings.

The second trial was carried out in 122 centers, most of which were quite small. In fact, the distribution of the center sizes was very similar to Distribution 1 from Figure 6. In this example, we focus on the missingness of the health score of each patient. The proportion of missing scores per center is shown in the right-hand-side panel of Figure 11. The overall proportion of the missing health scores was very low ($\hat{p} = 0.0090$, with $\hat{p}_w = 0.0012$) and most centers had no missing values at all. Only two centers had missing values: one center of size 10 (1 missing value) and one center of size 2 (both values missing). This is a very challenging case and the model adjustment proved helpful here. The model estimated from all centers yielded a large overdispersion estimate

($\hat{\rho}$ = 0.77), which was reduced to $\hat{\rho}$ = 0.50 by adjusting the α parameter. It is difficult to say that this model is a good approximation of the underlying null model, but the adjustment improved detection, in the sense that it made the *p*-value for the center of size 2 more significant (from *p*=0.015 when a standard beta-binomial was used to *p*=0.0027 after the proposed adjustment). Even though the center would be considered a statistically significant outlier by both approaches, having *p*-values that adequately quantify the extremeness (in terms of order of magnitude) of such outlying centers is crucial in CSM, where an overall score is computed based on the *p*-values of a battery of statistical tests [1].

5 Discussion

In a previous paper, we showed that linear mixed-effects models can be used to detect centers with atypical data on a continuous scale [17]. In the present paper, we extend the concept to the detection of centers with atypical event proportions, where the center size need not necessarily be the number of patients.

The proposed procedure can be applied to monitor any aspect of data quality that can be expressed as a binary event probability like, for instance, the proportion of missing values, the proportion of visit dates that fall on a Sunday, the proportion of untoward events, etc. The event of interest may also be derived from more complex data structures, like the transition probabilities between two states in a sequence of repeated binary observations. Note that the procedure is not appropriate for counts, e.g. for adverse events where the number of episodes is important. For such situations a modeling based on the Poisson distribution, or similar, would be indicated.

The procedure uses the beta-binomial distribution for a reliable, versatile, and automatic detection of centers having atypical event proportions thanks to a model adjustment step and a number of diagnostic checks and measures to address convergence problems. Note that the procedure but does not pursue an unbiased estimation of the underlying null model.

The simulation study confirmed that the power for departures towards the opposite boundary is maintained for contamination rates of interest (typically up to 5%), irrespectively of the values of μ_0 and μ_1. The adjustment does not decrease specificity, which was shown to be conservatively controlled across all simulated scenarios.

The procedure is robust with respect to unbalanced center sample sizes, but performance may obviously decline with decreasing sample sizes. However, the power to detect departures to the opposite boundary decreases mostly in scenarios where μ_0 is close to 0.5 and not when it is close to 0 or 1, the latter being of more interest in practice.

Regarding the assumptions of the procedure, it is fair to assume that the atypical phenomena are the exception, hence the small contamination rate, and that the variability of the typical centers is limited since subjects are comparable and treated as per protocol. However it is not unreasonable to imagine multiple atypical centers corresponding to different location parameters. Through simulation we investigated the power for detection of one or a few atypical center(s) of two classes with different location parameters ($\mu_1 \neq \mu_2$) in a population of typical centers (mean μ_0). Unsurprisingly, the power for detecting the μ_1-class is hardly affected when μ_2 is similar to μ_0 or μ_1, and is most affected when μ_2 is far away from μ_0, especially towards the opposite boundary, as this leads to a large estimated overdispersion in the hybrid model. In such cases however, while we may not detect centers of the μ_1-class, we would detect those of the μ_2-class. It can then be argued that an a posteriori investigation would also reveal the presence of the former, e.g. after removing the latter.

The setup of the detection procedure may raise the issue of multiple testing and the appropriateness of $\alpha_{crit} = 0.05$ in the decision rule. In this respect it is important to note that the purpose of the test procedure is not to detect atypical centers on the basis of multiple hypotheses (one per center), but rather to label centers as potentially atypical or not. The p-value is crucial in the sense that it allows to quantify the compatibility between the data from a given center and the parametric model derived from all the centers. Smaller p-values correspond to larger signals and have greater influence on the data inconsistency score for the centers concerned. Since in the computation of the data inconsistency score p-values from different test procedures are combined, the issue of multiple testing is appropriately handled at that level. The interpretation of flagging centers in Section 2 is a pragmatic way to allow definition of the performance criteria and the chosen value $\alpha_{crit} = 0.05$ corresponds to the mean fraction of false positives under the null-hypothesis (regardless of the total number of centers).

It is worth adding that the parameter γ in the adjustment step allows for some flexibility in the detectable level of contamination, as the adjustment is applicable for $\hat{p}$ in the $(0, \gamma)$ or $(1 - \gamma, 1)$ interval. Thus, for example, with $\gamma = 0.2$, the correction will stop at 20% contamination, but will have sufficient effect to guarantee the power at 10% contamination. The magnitude of the adjustment can also be adapted in terms of the target parameter β_{target} (or α_{target}) used in the interpolation. With values larger than 1 the adjustment becomes more aggressive (better power and smaller p-values), but values smaller than 1 may be considered to obtain a weaker adjustment. Clearly, the optimal parameter choices will depend on the specifications of the intended application.

Given the flexibility of the proposed procedure, and the satisfactory performance shown through simulations, it has potential as a building block in statistical monitoring applications, where the p-values from different detection procedures are combined to assign an overall score and a rank to each center [1]. We have implemented this approach in the SMART$^{\text{TM}}$software (Statistical Monitoring Applied to Research Trials), which can be used to identify data quality and consistency issues in multicentre clinical trials [18]. Reference [19] provides details on data quality checks performed in a large trial for patients with gastric cancer.

A Beta-binomial and related distributions

We review some useful results on relevant distributions: beta (continuous), binomial and beta-binomial (discrete).

Parametrisation. The shape parameters α and β are the natural parameters for the beta density, while the usual parameters in the context of the beta-binomial model are μ (mean) and ρ (overdispersion). Note that $0 < \mu < 1$ (success probability 0 or 1 is excluded) and $0 < \rho < 1$ (binomial is limiting case of beta-binomial when ρ tends to 0). These constraints are equivalent to $\alpha > 0$ and $\beta > 0$.

Conversion formulas.
$\rho = \frac{1}{1+\alpha+\beta}, \; \mu = \frac{\alpha}{\alpha+\beta}$ and $\alpha = \mu(\frac{1}{\rho} - 1), \; \beta = (1 - \mu)(\frac{1}{\rho} - 1)$.

Notation, mean and variance expressions.

	Beta(α, β) or (μ, ρ)	Beta-binomial(n, α, β) or (n, μ, ρ)	Binomial(n, p)
Mean	$\frac{\alpha}{\alpha+\beta} = \mu$	$\frac{n\alpha}{\alpha+\beta} = n\mu$	np
Variance	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \mu(1-\mu)\rho$	$\frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} = n\mu(1-\mu)(1+(n-1)\rho)$	$np(1-p)$

Symmetry properties. The beta pdf is given by $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ for $0 < x < 1$ where B is the beta function. Exchanging α and β corresponds to exchanging arguments x and $1 - x$.

Event counts x follow a Beta-binomial(n, μ, ρ) model iff non-event counts $n - x$ follow a Beta-binomial($n, 1 - \mu, \rho$) model (this is just a generalisation of a similar property of the binomial). As a consequence, we can limit our discussion to the case $\mu \leq 0.5$.

Shape of the beta density. By computing derivatives one can verify the behaviour shown below at the left boundary (symmetric result at the right boundary in terms of β).

	$\lim_{x\to0+} f(x)$	$\lim_{x\to0+} \frac{df}{dx}$	$\lim_{x\to0+} \frac{d^2f}{dx^2}$
$\alpha < 1$	$+\infty$	$+\infty$	$+\infty$
$\alpha = 1$	c	$+\infty$	$+\infty$
$1 < \alpha < 2$	0	$+\infty$	$+\infty$
$\alpha = 2$	0	c	$+\infty$
$2 < \alpha < 3$	0	0	$+\infty$
$\alpha = 3$	0	0	c
$\alpha > 3$	0	0	0

For $\alpha > 2$ and $\beta > 2$, the density is unimodal with sigmoidal tails. If $\alpha < 2$ and $\beta < 2$ there is a pole on either side and the density is U-shaped. If one is large (> 2) and the other is small (< 1) we have a pole on one side and a tail on the other side. In this paper we model event rates with a small amount of overdispersion, say at most 0.1, and we refer to Figure 12 to get an idea of the variance in such models. The densities are unimodal or L-shaped, but never U-shaped.

References

[1] Venet D, Doffagne E, Burzykowski T, Beckers F, Tellier Y, Genevois-Marlin E, Becker U, Bee V, Wilson V, Legrand C, Buyse M. A statistical approach to central monitoring of data quality in clinical trials. *Clin Trials* 2012; 9: 705–713.

[2] Kirkwood AA, Cox T, Hackshaw A. Application of methods for central statistical monitoring in clinical trials. *Clin Trials* 2013; 10: 783–806.

[3] Pogue JM, Devereaux PJ, Thorlund K, Yusuf S. Central statistical monitoring: detecting fraud in clinical trials. *Clin Trials* 2013; 10: 225–35.

[4] Lindblad AS, Manukyan Z, Purohit-Sheth T, Gensler G, Okwesili P, Meeker-O'Connell A, Ball L, Marler JR. Central site monitoring: results from a test of accuracy in identifying trials and sites failing Food and Drug Administration inspection. *Clin Trials* 2014; 11: 205–17.

[5] Chuang-Stein C. An Application of the beta-binomial model to combine and monitor medical event rates in clinical trials. *Drug Information Journal* 1993; 27: 515.

[6] Young-Chu Y, Chan K. Pooling overdispersed binomial data to estimate event rate. *BMC Medical Research Methodology* 2008; 8: 58.

[7] Griffiths DA. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics* 1973; 29(4): 637–648.

[8] Williams DA. The analysis of binary responses from toxicological experiments involving reproduction and teratogenecity. *Biometrics* 1975; 31(4): 949–952.

[9] Kulinskaya E. On two-sided *p*-values for non-symmetric distributions. arXiv:0810.2124v1, 2008.

[10] Tripathi RC, Gupta RC and Gurland J. Estimation of parameters in the beta binomial model. *Ann Inst Statist Math* 1994; 46(2): 317–331.

[11] Kleinman JC. Proportions with extraneous variance: single and independent samples. *J Am Stat Assoc* 1973; 68: 46–54.

[12] R Development Core Team. 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/, 2011.

[13] Yee TW. VGAM: Vector Generalized Linear and Additive Models. R package version 0.8-7. URL http://CRAN.R-project.org/package=VGAM, 2012.

[14] Yee TM. VGAM Family Functions for Univariate Distributions. https://www.stat.auckland.ac.nz/ yee/VGAM/doc/univar.pdf, 2004.

[15] Tarone RE, Testing the goodness of fit of the binomial distributions. *Biometrika* 1979; 66(3): 585–590.

[16] Bloomfield P. *Fourier Analysis of Time Series, 2nd Edition*. John Wiley, 2000.

[17] Desmet L, Venet D, Doffagne E, Timmermans C, Burzykowski T, Legrand C, Buyse M. Linear mixed-effects models for central statistical monitoring of multicenter trials. *Statist Med* 2014; 33(30): 5265–5279.

[18] Timmermans C, Venet D, Burzykowski T. Data-driven risk identification in phase III clinical trials using central statistical monitoring. *Int J Clin Oncol* 2015, DOI 10.1007/s10147-015-0877-5.

[19] Timmermans C, Doffagne E, Desmet L, Venet D, Legrand C, Burzykowski T, Buyse M. Using central statistical monitoring to assess data quality and consistency in the Stomach cancer Adjuvant Multi-Institutional group Trial (SAMIT). *Gastric Cancer* 2015, DOI 10.1007/s10120-015-0533-9.

Table 1: Setup for simulations in the balanced case

number of centers	$N = 100$
number of atypical centers	$N_1 = 1, 2, 5, 10, 20, 40, 50$
all center sizes	$n_i = n = 100$ (balanced)
null mean	$\mu_0 = 10^{-6}, 10^{-4}, 0.001, 0.01, 0.1, 0.2, 0.4, 0.5$
alternative mean	$\mu_1 = 10^{-6}, 10^{-4}, 0.001, 0.01, 0.04, 0.08, ..., 0.98, 0.99, 0.999, 1 - 10^{-4}, 1 - 10^{-6}$
overdispersion level	$\rho = 0$ (binomial),0.01 (mild overdispersion), 0.1 (large overdispersion)
number of replications	N_{sim} adjusted s.t. $N_{sim}N_1 = 2000$

Table 2: Setup for simulations in the unbalanced case

number of centers	$N = 20$
number of atypical centers	$N_1 = 1, 4, 10$
sizes (typical centers)	random drawn from size distributions (unbalanced)
sizes (atypical centers)	random, fixed at 5% quantile or fixed at 50% quantile
null mean	$\mu_0 = 10^{-6}, 10^{-4}, 0.001, 0.01, 0.05, ..., 0.45, 0.5$
alternative mean	$\mu_1 = 10^{-6}, 10^{-4}, 0.001, 0.01, 0.02, ..., 0.98, 0.99, 0.999, 1 - 10^{-4}, 1 - 10^{-6}$
overdispersion level	$\rho = 0, 0.01, 0.1$
number of replications	N_{sim} adjusted s.t. ($N_{sim}N_1 = 2000$)

Table 3: Specificity in the simulations

	Min.	1%-tile	1st Qu.	Median	(Mean)	3rd Qu.	Max.
Table 1	0.9460	0.9562	0.9913	1.0000	(0.9931)	1.0000	1
Table 2, Distribution 2, all random	0.9572	0.9587	0.9782	0.9942	(0.9881)	0.9993	1
Table 2, Distribution 1, small atypical sample size	0.9290	0.9577	0.9779	0.9825	(0.9848)	0.9935	1
Table 2, Distribution 1, median atypical sample size	0.9510	0.9585	0.9750	0.9945	(0.9877)	1.0000	1
Table 2, Distribution 2, small atypical sample size	0.9555	0.9690	0.9860	0.9900	(0.9909)	0.9982	1
Table 2, Distribution 2, median atypical sample size	0.9680	0.9741	0.9843	0.9950	(0.9918)	1.0000	1
Table 2, Distribution 3, small atypical sample size	0.9555	0.9594	0.9831	0.9990	(0.9909)	1.0000	1
Table 2, Distribution 3, median atypical sample size	0.9750	0.9810	0.9886	0.9953	(0.9939)	1.0000	1

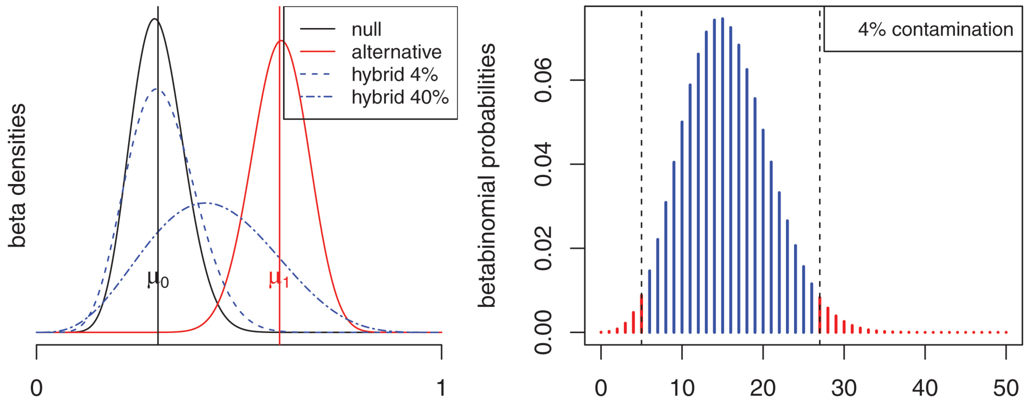


Figure 1: Left panel: null, alternative, and estimated hybrid densities for two levels of contamination. Right panel: hybrid beta-binomial distribution for a center of size 50 (critical sizes delimited with dashed line). Parameters are $\mu_0 = 0.3$ and $\mu_1 = 0.6$; $\rho_0 = \rho_1 = 0.02$.

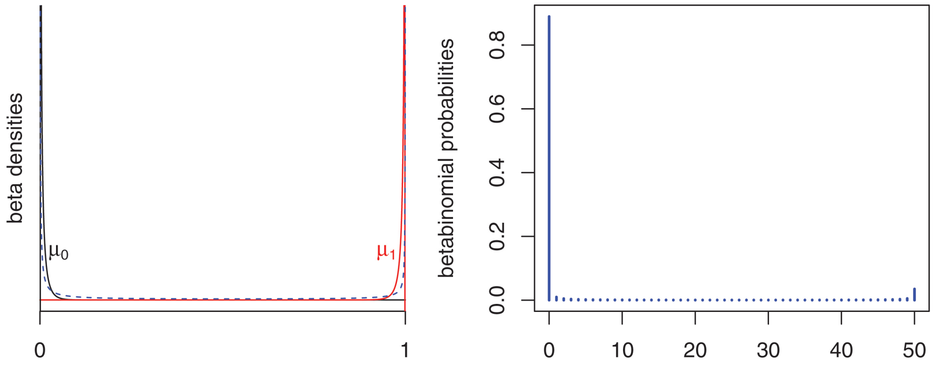


Figure 2: Same as Figure 1, but with $\mu_0 = 0.001$ and $\mu_1 = 0.999$; $\rho_0 = \rho_1 = 0.02$ (4% contamination). In the hybrid distribution all p-values exceed $\alpha = 0.05$.

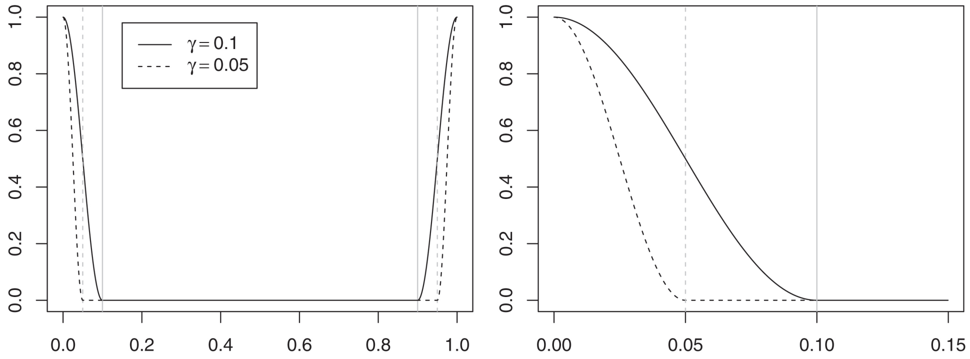


Figure 3: Left panel: plot of the function λ on $[0, 1]$, for two values of the parameter γ. Right panel: zoom-in on the left boundary.

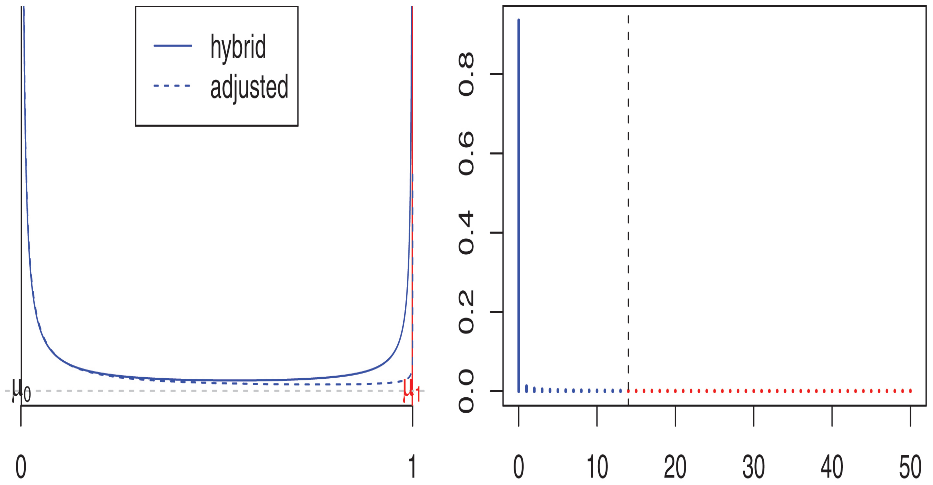


Figure 4: Adjustment in the example of Figure 2. Left panel: hybrid density and adjusted density. Right panel: critical region under the adjusted distribution.

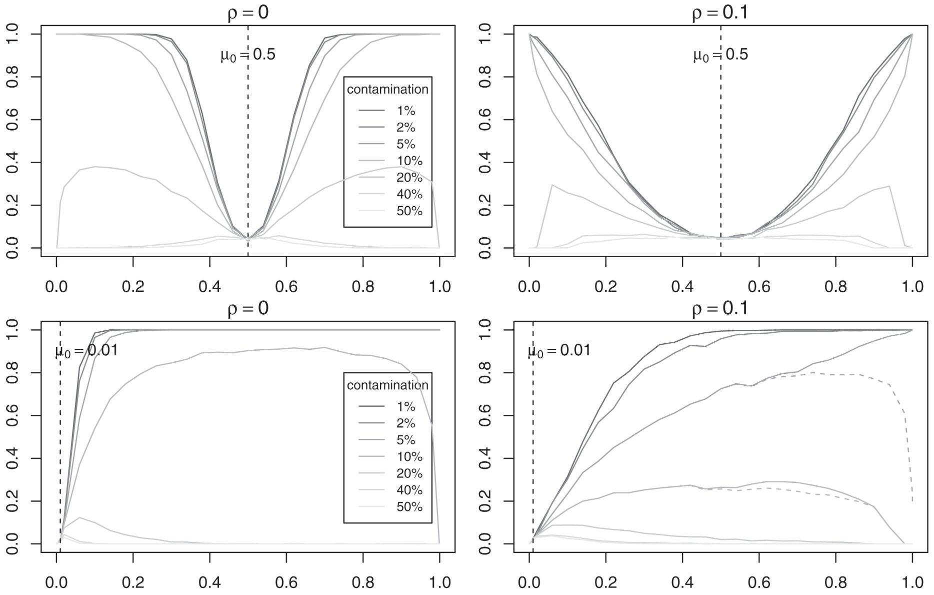


Figure 5: Power *versus* μ_1 for fixed μ_0 (indicated by dashed vertical lines) and $\rho = 0$ (left) and 0.1 (right).

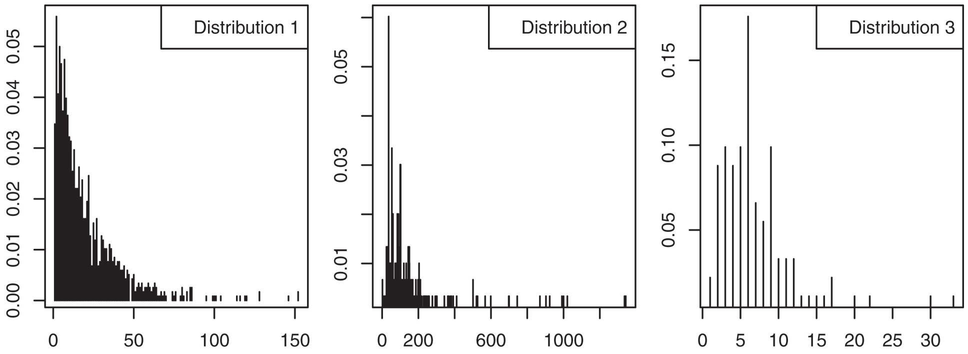


Figure 6: Probability distribution functions based on observed center sizes in three actual clinical trials.

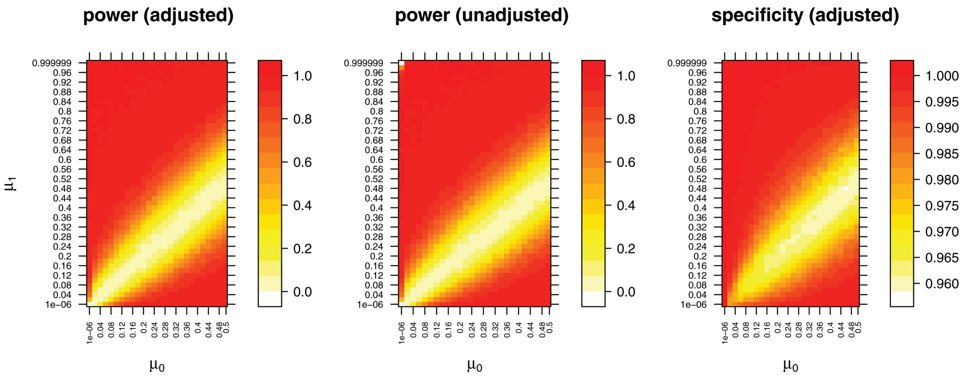


Figure 7: Performance in the unbalanced case with Distribution 2. Heatmaps on a fine gridsize (0.02) for $\rho = 0.01$ and 5% contamination.

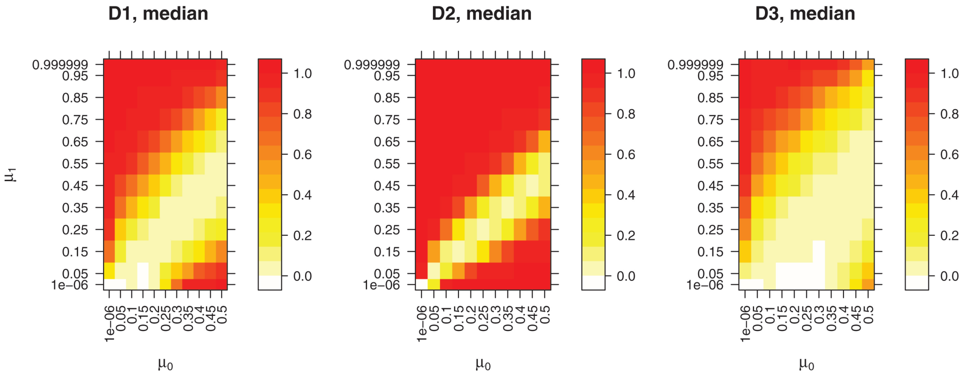


Figure 8: Power levels on a grid of (μ_0, μ_1) values with 5% contamination and $\rho = 0.01$. Typical-centers sample sizes were drawn at random from Distribution 1 (left panel), Distribution 2 (middle panel), or Distribution 3 (right panel). Atypical-center sample sizes were fixed at the median of the corresponding sample size distribution (13 for Distribution 1, 100 for Distribution 2, and 6 for Distribution 3).

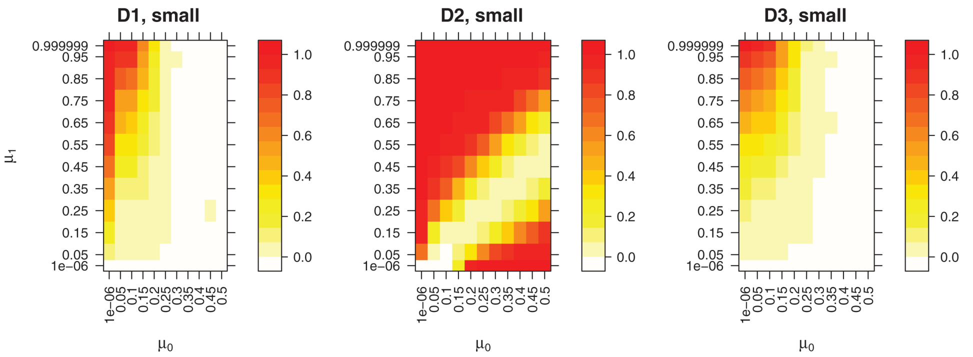


Figure 9: Power levels on a grid of (μ_0, μ_1) values with 5% contamination and $\rho = 0.01$. Typical-centers sample sizes were drawn at random from Distribution 1 (left panel), Distribution 2 (middle panel), or Distribution 3 (right panel). Atypical-center sample sizes were fixed at the 5%-tile of the corresponding sample size distribution (2 for Distribution 1, 24 for Distribution 2, and 2 for Distribution 3).

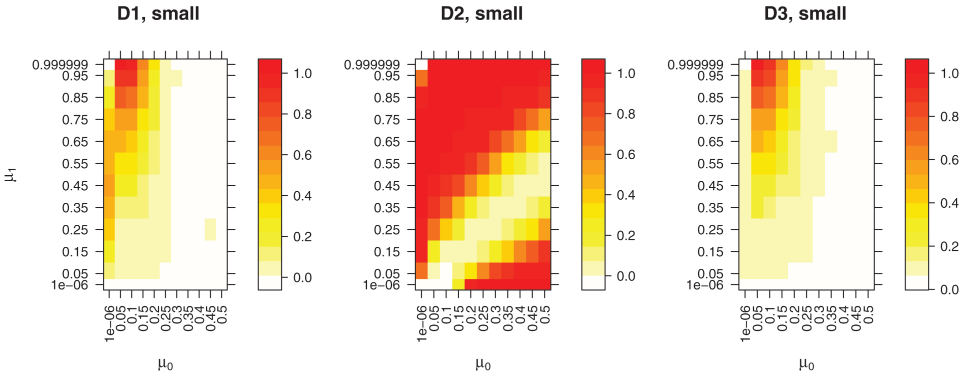


Figure 10: Same as Figure 9 but for the (unadjusted) beta-binomial-model procedure.

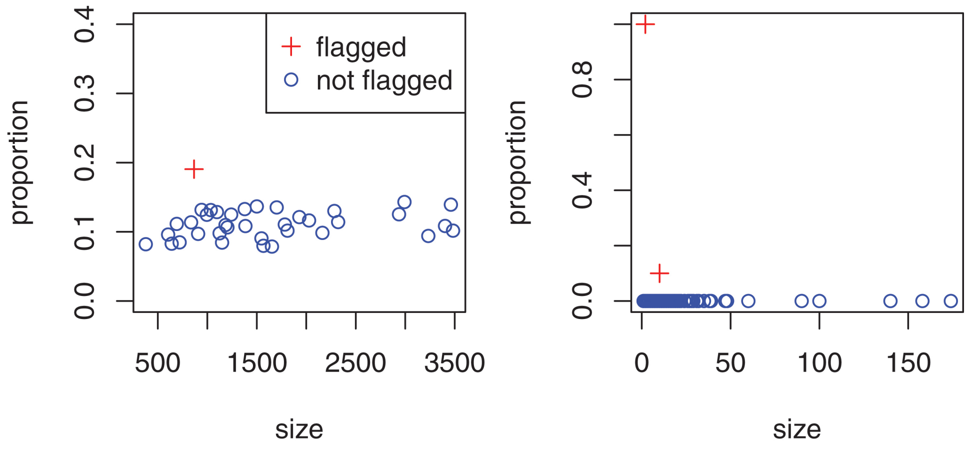


Figure 11: Scatter plots of sizes (x-axis) and observed proportions (y-axis) in Example 1 (left panel) and Example 2 (right panel). Note that in Example 1 the sizes are not the center sizes as multiple visits per patient are aggregated.

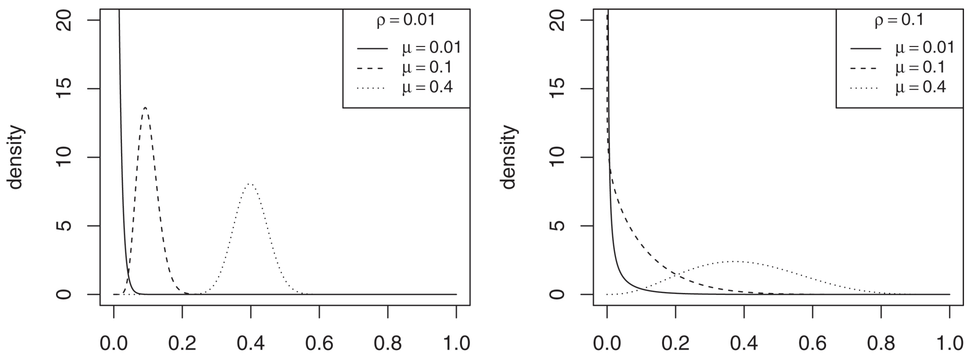


Figure 12: Beta densities for two levels of overdispersion.