Special Issue Paper

Statistics in Medicine

Received 31 October 2013, Accepted 15 August 2014 Published online 12 September 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6294

Linear mixed-effects models for central statistical monitoring of multicenter clinical trials

L. Desmet,[a*†] D. Venet,[b] E. Doffagne,[c] C. Timmermans,[c] T. Burzykowski,[c,d] C. Legrand[a] and M. Buyse[d,e]

Multicenter studies are widely used to meet accrual targets in clinical trials. Clinical data monitoring is required to ensure the quality and validity of the data gathered across centers. One approach to this end is central statistical monitoring, which aims at detecting atypical patterns in the data by means of statistical methods. In this context, we consider the simple case of a continuous variable, and we propose a detection procedure based on a linear mixed-effects model to detect location differences between each center and all other centers. We describe the performance of the procedure as a function of contamination rate and signal-to-noise ratio. We investigate the effect of center size and variance structure and illustrate the use of the procedure using data from two multicenter clinical trials. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: multicenter clinical trial; statistical monitoring; error detection; contamination rate; signal-to-noise ratio; linear mixed-effects model

1. Introduction

Monitoring of clinical trials is a crucial process to ensure not only the safety and well-being of the patients but also the validity of the trial outcomes and conclusions. Current practices rely mostly on source data verification (SDV) to verify that data are genuine, accurate, and valid. As SDV typically involves on-site visits, the costs of this approach are huge and represent a very substantial part of the budget of a clinical trial [1]. In addition, on-site verifications do not allow a comparison of data from multiple centers. With central statistical monitoring (CSM), the idea is to gather information across centers and to look at the 'big picture'.

The idea of using statistical methods for clinical data monitoring has been around for some time. For an overview and discussion of the use of biostatistical methods in the detection and prevention of fraud, refer to [2]. Actual use of these methods has however been quite limited until now, which is not surprising for an industry that has for a long time considered on-site monitoring and SDV as the norm. The situation has recently evolved, however, with CSM receiving increased attention as a cost-effective way to improve data quality. Both the US Food and Drug Administration and the European Medicines Agency encourage a switch from current practices to centralized and risk-based monitoring [3, 4].

Nowadays, with the growing adoption of electronic data capture systems in clinical trials, the data are centrally available in near real time. It is therefore possible to apply a battery of statistical tests on the totality of the data already available at any time during the trial. We have described elsewhere a system that performs a large number of tests on all variables collected and combines the resulting information

[a] *Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain, Louvain-la-Neuve, Belgium*
[b] *Institut de Recherches Interdisciplinaires et de Développements en Intelligence Artificielle (IRIDIA), Université Libre de Bruxelles, Brussels, Belgium*
[c] *International Drug Development Institute (IDDI) S.A., Louvain-la-Neuve, Belgium*
[d] *Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Diepenbeek, Belgium*
[e] *International Drug Development Institute (IDDI) Inc., San Francisco, CA, U.S.A.*
**Correspondence to: L. Desmet, ISBA, Université catholique de Louvain, Voie du Roman Pays 20, bte L1.04.01 B-1348 Louvain-la-Neuve, Belgium.*
[†] *E-mail: lieven.desmet@uclouvain.be*

in an overall assessment of each center as compared with all other centers [5]. The key idea behind this approach is to compare the data from one center with the empirical distribution of the data from all other centers. For each variable, the compatibility of a center with the general tendency across all other centers is quantified by the p-values of the test procedures that can be carried out given the type and structure of the variable as implied by the study protocol.

The aim of this paper is to describe one of the statistical test procedures used to compare each center with all other centers, when interest focuses on the location parameter of a continuous variable. In Section 2, we introduce a simple test procedure for univariate normal data and derive its performance for different signal-to-noise ratios (SNRs) and contamination rates. This provides a prototype for the test procedure described in Section 3 where the hierarchical structure of the data in a typical clinical trial (measures taken on patients treated at specific centers) is accounted for through a linear mixed-effects model. Section 4 is devoted to assessing the performance of the procedure in terms of sensitivity (or power) and specificity through simulation studies. Section 5 concludes with examples using data from two actual clinical trials.

2. Detecting a shift in location for normal data

In this section, we consider observations for some normally distributed variable. If all observations agree with a single normal model, with a plausible *location* parameter for the variable at hand, we have no reason to label any of the observed values as atypical. Suppose now that some of the observations come from a model with a different location. Then, we wish to detect these values as being different, especially if the shift is substantial.

2.1. Formal problem and detection procedure

From a methodological point of view, we will consider that the data are a mixture where the majority of the observations correspond to a certain normal model, the *null model*, while a few observations agree with a shifted version of this model, the *alternative model*. The former observations can be regarded as the *clean data*, the latter as the *contaminants*.

To fix notations, assume we have a mixed sample of size $n = n_0 + n_1$ where

- n_0 *i.i.d.* observations (the clean data) are distributed as $N(\mu_0, \sigma^2)$ (the null model) and
- n_1 *i.i.d.* observations (the contaminants) are distributed as $N(\mu_1, \sigma^2)$ (the alternative model), with $\mu_1 \neq \mu_0$.

This setup is very similar to the one introduced by Guttman [6] and often referred to as the *location-shift model*. Note that n_1 may be 0 and will typically be small compared with n_0 (in any case $n_1 < n_0$ to avoid problems of identifiability).

The general problem is thus to detect the contamination, if present, in the combined sample (*hybrid sample*) as defined earlier, but without a priori information about the parameters. In case no contamination is present ($n_1 = 0$), the null model is directly estimated from all observations and can be used to assess the plausibility of any value. If some atypical observations are present, we cannot obtain the null model from all observations, but we may assume that the normal model estimated from all data, the so-called hybrid model $N\left(\mu_{hybrid}, \sigma^2_{hybrid}\right)$, is a good approximation of the null model, provided that the contamination rate is sufficiently small.

In that case, we may also assume that the contaminant observations are eccentric with respect to the clean data and they will tend to cluster in one of the tails of the null model if the location shift is substantial. Detection of the contaminants can thus be performed by looking at the tails delimited by the $\alpha/2$ and $1-\alpha/2$ quantiles for some small value α. One can consider any observation falling in the tails as potentially suspicious, at least if more of them are observed than expected given α and the sample size.

To formalize this, we associate with each observation x a probability $p(x)$ that reflects how extreme the observation is with respect to the hybrid distribution:

$$p(x) = \begin{cases} 2P\{X < x\} & \text{if} \quad x \leqslant \mu_{hybrid} \\ 2P\{X > x\} & \text{if} \quad x > \mu_{hybrid} \end{cases}, \quad X \sim N\left(\mu_{hybrid}, \sigma^2_{hybrid}\right)$$

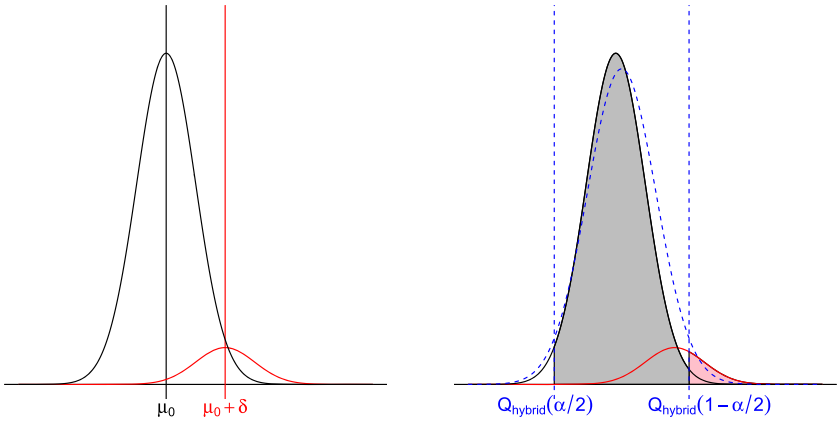

Figure 1. Left panel: densities of the null and the alternative models, each scaled according to their weight in the mixture (90% and 10%, respectively, in this example). Right panel: same, with hybrid density added (dashed line); the vertical lines are the 0.025 and 0.975 quantiles with respect to this density; power and specificity are the shaded areas under the alternative density curve and the null density curve, respectively.

This is very similar to a p-value in the interpretation of a t-test (two-sided, hence the factor 2). It is natural to *flag an observation x as suspicious* if and only if $p(x) < \alpha$.

With this decision rule, the type I error probability under the null model is equal to α (hence, we will refer to α as the *significance level*). The densities from the different distributions involved are illustrated in Figure 1. Our detection procedure implicitly defines a detection region and a nondetection region, based on the $\alpha/2$ and the $1-\alpha/2$ quantiles of the hybrid distribution (vertical lines in the right panel of Figure 1).

The simple detection procedure can be summarized as follows:

- Estimation step: estimate the hybrid model from all data.
- Evaluation step: assign a p-value to each observation.
- Detection step: flag observations according to the decision rule.

2.2. *Theoretical performance of the procedure*

The notion of performance is based on two desirable properties of the detection procedure:

- Sensitivity (or power): the ability of the procedure to flag (detect) the contaminant observations.
- Specificity: the ability of the procedure to avoid flagging clean observations.

We adopt the usual terminology of *false positives* for falsely detected observations (compare with the type I error in a hypothesis test) and *false negatives* for observations that should have been detected but were not.

The advantage of our simple problem setup is that we can compute the power and specificity theoretically (as probabilities). If we write $\mu_1 = \mu_0 + \delta$ and $w = \frac{n_1}{n_0+n_1}$, where w is the contamination rate, we can derive the probabilities for a false positive outcome and for a false negative outcome, respectively, as

$$p_{FP} = F_Z\left\{\frac{\delta}{\sigma}w + \sqrt{1+\left(\frac{\delta}{\sigma}\right)^2 w(1-w)}\,Q_Z\left(\frac{\alpha}{2}\right)\right\} + 1 - F_Z\left\{\frac{\delta}{\sigma}w - \sqrt{1+\left(\frac{\delta}{\sigma}\right)^2 w(1-w)}\,Q_Z\left(\frac{\alpha}{2}\right)\right\} \text{ and} \tag{1}$$

$$p_{FN} = F_Z\left\{\frac{\delta}{\sigma}(w-1) - \sqrt{1+\left(\frac{\delta}{\sigma}\right)^2 w(1-w)}\,Q_Z\left(\frac{\alpha}{2}\right)\right\} - F_Z\left\{\frac{\delta}{\sigma}(w-1) + \sqrt{1+\left(\frac{\delta}{\sigma}\right)^2 w(1-w)}\,Q_Z\left(\frac{\alpha}{2}\right)\right\}, \tag{2}$$

where F_Z and Q_Z refer, respectively, to the distribution function and the quantile function of the standard normal distribution (details in Appendix).

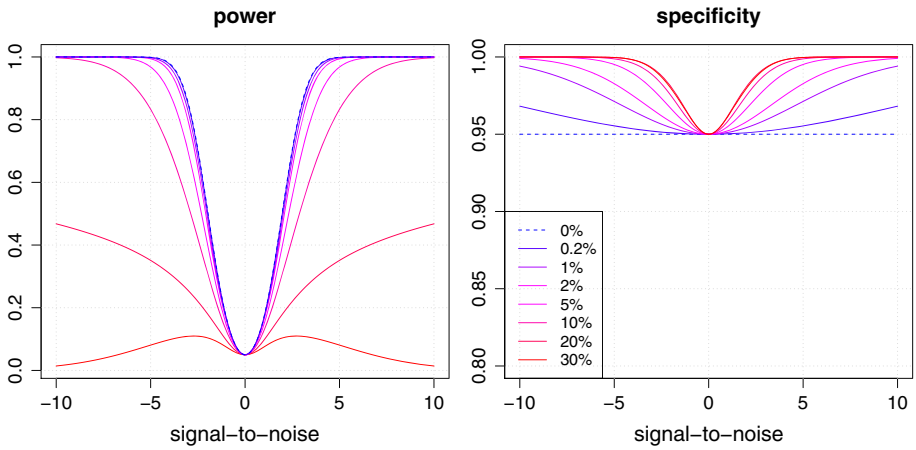

Figure 2. Power (left panel) and specificity (right panel) as functions of the signal-to-noise ratio. Curves are drawn for different values of the contamination rate $w = 0\%, 0.2\%, 1\%, 2\%, 5\%, 10\%, 20\%, 30\%$ (color scale: blue to red) and for a significance level (α) of 5%.

In the aforementioned probabilities, the parameters δ and σ always appear together as the ratio δ/σ, which is the *SNR*. The performance of the procedure directly follows from $power = 1 - p_{FN}$ and $specificity = 1 - p_{FP}$, which are functions of *SNR* and w for a fixed significance level α (also illustrated in Figure 1). The formulae still work with $w = 0$ (in which case the hybrid model coincides with the null model) and describe the limiting performance when w tends to 0, that is, the performance to be expected for the detection of an infinitely small contamination rate. Figure 2 shows the power and specificity as functions of *SNR*, for different levels of the contamination rate w (bundle of curves).

The curves are symmetric, in line with the symmetry of the normal model (therefore, in the remainder of this section, we will assume tacitly that *SNR* is positive). They also inherit the smoothness properties present in F_Z. Regardless of the contamination, the power reaches its (local) minimum of α for arbitrarily small *SNR*. Note that when $\delta = 0$, all observations detected are false positives under the null (specificity of $1 - \alpha$), and the limiting value for the power with arbitrarily small *SNR* is α (in the limit, the null model, the alternative model, and the hybrid model all coincide). For small levels of contamination, say up to 10%, the power increases monotonically for increasing values of *SNR* and rapidly reaches 100%. For larger levels of contamination, the detection is more difficult, hence the smaller slope of the power curves. Beyond a certain level of contamination, the monotonic behavior is lost, and the power curve tends to zero for large *SNR* values (e.g., curve for $w = 30\%$). The reason is that a large amount of contamination influences the hybrid model so much that contaminants end up in the central region rather than in the tails of the distribution. This is a desirable and necessary property for the purpose of detecting small amounts of contamination (numerical investigations indicated that the sudden change in the power function occurred for a value of the contamination rate $w \approx 20.65\%$: with less contamination, the power tends to 1 while with more contamination, the power tends to 0, each time for increasing *SNR*).

The type I error probability of the procedure is conservatively controlled. The specificity is equal to $1 - \alpha$ for $SNR = 0$ or $w = 0$ (case of no contamination), and greater than $1 - \alpha$ otherwise, with the curve tending to 1 as the *SNR* increases. The effect of the contamination rate on the specificity curves is in the opposite direction as compared with power curves. The curve is flat when there is little contamination, but it gets progressively steeper for increasing levels of contamination.

We note that a normal model may not always be adequate for the hybrid data at hand if contamination is truly present (the normality hypothesis may be rejected with standard tests, and bimodality can be obvious). However, our primary goal is not to describe the distribution but to detect atypical observations.

2.3. Alternative detection procedures

The detection procedure introduced in Section 2 is based on a hybrid model estimated from all data, which differs from the null model in the presence of contaminated data. It might seem appropriate to obtain a more accurate estimate of the null model by using estimators that are less influenced by the atypical observations.

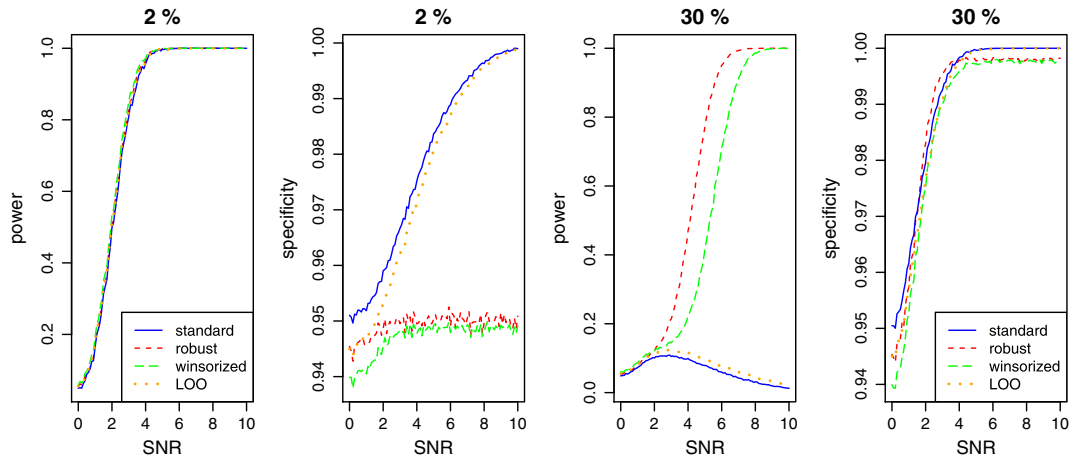

Figure 3. Comparison of four procedures. Power and specificity at 2% and 30% contamination.

In this section, we use simulations to study the performance of four detection procedures: (1) the procedure described earlier using the hybrid model, called here the standard procedure; (2) a robust procedure in which the hybrid model is estimated from all data with robust measures median and median absolute deviation; (3) another robust procedure in which the hybrid model is estimated from *winsorized* data (obtained by shrinking outlying observations to the border of the main part of the data); and (4) a leave-one-out (LOO) procedure in which each observation is compared with its proper reference model, namely, the model fitted to all data except the observation itself.

Simulations were performed using hybrid datasets of size 100, generated at random according to different levels of contamination; 1000 replications were performed for each simulated scenario.

The graphs in Figure 3 present the results of the simulation study. They show that robust procedures are far more sensitive than both the standard procedure and the LOO procedure for large levels of contamination (such as 30%). However, for the low levels of contamination generally of interest in the present context (such as 2%), the gain in power is negligible and the robust procedures induce a serious loss of specificity; that is, they have the undesirable property of flagging as atypical centers that are not. We also note that the LOO procedure does not perform better than the standard procedure and is computationally more costly. All in all, the standard procedure is fit for our purpose, and we therefore use it in the remainder of this paper.

Many authors have used the framework of *finite mixture modeling* to detect atypical observations as a separate component. Everitt and Hand [7] discuss the problem of parameter estimation in a mixture of normal components. However, applicable methods assume that the number of components (and thus the existence of contamination) is known beforehand and may involve computationally complex algorithms.

In the same context of mixtures, Guttman [6] introduced the location-shift model and considered the problem of detecting a single outlier in a normal population, relying on a Bayesian approach. For a recent overview of the field and applications in cluster analysis and outlier detection, we refer to Früwirth-Schnatter [8].

3. A testing procedure for center means in clinical data

We now turn to the setting of a multicenter clinical trial where several patients are enrolled per center. For simplicity, we assume that we have one measurement per patient, although the methodology can be extended to repeated measurements for each patient. The idea is again to detect atypical centers, and we do this by looking at the center means and assessing their position with respect to the other data, using a suitable model for all data. The linear mixed-effects model has been widely used for longitudinal or clustered data and is a natural framework in which the random center effect accounts for the variation across centers [9, 10].

In this setup, values of the variable of interest are denoted y_{ij} where i is the center index ($i = 1, \ldots, N_c$) and j is the patient index ($j = 1, \ldots, N_i$). We can write

$$y_{ij} = \mu + \gamma_i + \varepsilon_{ij},$$

where μ is the fixed effect, γ_i is the random center effect, and ε_{ij} is the random residual error (corresponding to individual patients). The γ_i are *i.i.d.* $N\left(0, \sigma_c^2\right)$, independent from the ε_{ij} that are *i.i.d.* $N\left(0, \sigma_r^2\right)$. We will here focus on this simple random effects model, but fixed effects can be added to the model if appropriate, for instance, to account for repeated visits of the patients over time.

The mean value for each center is defined as $\bar{y}_{i\bullet} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$ (for the i-th center). Under the assumptions of the model, the $\bar{y}_{i\bullet}$ considered as random variables are distributed as $N\left(\mu, \sigma_c^2 + \frac{\sigma_r^2}{N_i}\right)$. Therefore, it is natural to look at the quantities $\bar{y}_{i\bullet} - \mu$ and use an $N\left(0, \sigma_c^2 + \frac{\sigma_r^2}{N_i}\right)$ reference distribution to assess to what extent centers are atypical. The rationale for detection is the following: if one or more centers have a different location (say with a fixed-effect mean value $\mu + \delta$ instead of μ), the position of the corresponding $\bar{y}_{i\bullet} - \mu$ will also be shifted with respect to this reference distribution.

We can develop a testing procedure that is similar to the one for contaminated normal data. Again, we make the assumption that the data are a mixture of clean and contaminated observations in the sense that a minority of the centers may have a shifted location but the same within-center variance as the other centers. As before, this implies that the model fitted to all data is a hybrid model, and the null model parameters, which we denote μ, σ_c and σ_r, remain unknown. Instead, we obtain estimates $\hat{\mu}_{hybrid}$, $\hat{\sigma}_{c,hybrid}$ and $\hat{\sigma}_{r,hybrid}$ from the hybrid model fit and use these as input for the reference distribution. The evaluation of each center is then carried out as in the procedure for contaminated normal data.

The detection procedure for hierarchical data is summarized as follows:

- Estimation step: estimate hybrid model from all data, using the linear mixed-effects model.
- Evaluation step: assign a p-value to each center (using the $N\left(0, \hat{\sigma}^2_{c,hybrid} + \hat{\sigma}^2_{r,hybrid}/N_i\right)$ distribution for the $\bar{y}_i - \hat{\mu}_{hybrid}$).
- Detection step: flag centers according to the decision rule.

In addition to the *SNR* and contamination rate, the size of the centers plays a role in this procedure. However, in the particular case of balanced data, that is, when each center i has the same number of patients $N_i = N$, there is a direct link between the problem considered here and the one in Section 2. Indeed, in this case and under the null model assumptions, the center means all have a common variance $\sigma_c^2 + \frac{\sigma_r^2}{N}$ and form a normal population. Adding the same shift to all observations in some of the centers corresponds to adding that shift to the center means. This link will be further discussed in the numerical study.

4. Assessing the performance of the procedure

In this section, we assess the performance of the testing procedure developed in Section 3 by means of an extensive simulation study. In Section 4.2, we consider balanced centers, and we focus, as aforementioned, on the effects of the level of contamination and the *SNR*. We show that the performance can be predicted if the null model parameters are known (which they are not in most real situations). In Section 4.3, we look at the impact of unbalanced center sizes and the effect of the size of the atypical center(s). Finally, we discuss the interpretation of the performance and investigate what happens when the location-shift model assumptions are violated.

4.1. Simulation approach

Simulations were carried out by repeating the following steps:

- Generate a data set according to some scenario (group and parameter setup).
- Run the test procedure.
- Evaluate by comparing obtained and expected results (atypical centers are known).

In practice, all data are first generated from the null model. A number of atypical centers are then chosen (the contamination rate is the ratio of the number of atypical centers over the total number of centers), and their data are obtained by adding a shift δ (identical for all atypical centers) to the original observations.

The significance level α is 0.05 in all simulations reported in this section. The performance is computed by counting the number of true positives (#*TP*), true negatives (#*TN*), false positives (#*FP*), and false negatives (#*FN*) and evaluating

$$\text{power} = \frac{\#TP}{\#TP + \#FN}, \quad \text{specificity} = \frac{\#TN}{\#FP + \#TN}.$$

Note that, because we typically look at small contamination rates, we may have only a few atypical centers; however, by averaging over a sufficient number of replications, we can obtain reliable performance estimates. In the simulation studies, we perform a number N_{sim} of replications. In each replication, the observed power is based on counting the true positives among the $N_{atypical}$ atypical centers. The simulation exercise thus yields a sequence of N_{sim} *i.i.d.* random variables denoted X_k for $k = 1, \ldots, N_{sim}$. It is reasonable to assume that $X_k \sim$ Binomial$(N_{atypical}, \pi)$ where π is the power to be estimated. From the central limit theorem, the average $\hat{\pi}$ over simulations of the $X_k/N_{atypical}$ is approximately normal; therefore, we can take $\sqrt{\hat{\pi}\,(1-\hat{\pi})\,/(N_{atypical}N_{sim})}$ as the standard error of the estimated power. The number of atypical centers directly appears in the formula; hence, the precision for the power is better for larger amounts of contamination, for the same total number of centers. For this reason, the number of replications was adjusted to the level of contamination for all the simulation exercises in this section, resulting in a standard error for the estimated power of at most 0.011 (taking 0.5 for $\hat{\pi}$). Obviously, the precision of the specificity estimate is always higher.

4.2. Influence of contamination rate and signal-to-noise ratio in a balanced setup

The basic setup for the first set of simulations is outlined in Table I. Different rates of contamination are considered: 0.5%, 2%, 10%. Obviously, the magnitude of the signal depends on δ and not on the location of the null model, so we may take $\mu = 0$ without loss of generality. In the balanced case, an absolute shift δ corresponds to an effective *SNR* of $\delta/\sqrt{\sigma_c^2 + \frac{\sigma_r^2}{N}}$. Therefore, values for δ will be chosen to obtain a range of *SNR* from -10 to 10 (as in Figure 2) with sufficient resolution near zero to capture the curvature. The parameter σ_c is set to 1, again without loss of generality (changing this parameter corresponds to a rescaling of the absolute shifts). The choice of $\sigma_r = 4$ reflects the fact that the within-center variance (variance between individual patient observations) tends to be larger in most clinical trials than the variance between the centers. We will assess the impact of this parameter on performance in Section 4.3.

Average performance curves are presented in Figure 4 where the x-axis represents the absolute shift δ. There is a similarity with the characteristics for contaminated normal data. As expected, the power of the procedure increases with increasing absolute value of the shift, but increasing levels of contamination lead to a decreasing power. The specificity also behaves in the same way as in Figure 2 and is conservatively controlled.

In the balanced case, we can actually predict the performance, given the parameters of the scenario. Indeed, with *SNR* values calculated as $\delta/\sqrt{\sigma_c^2 + \frac{\sigma_r^2}{N}}$ (using the variance parameters of the null model) and taking into account the level of contamination, the theoretical values for power and specificity can be computed using expressions (1) and (2). These values (plotted in Figure 4 at the corresponding absolute shifts) are in good agreement with the average performance curves.

Though mathematically trivial in the balanced case, the effect of center size is worth discussing. For a fixed shift δ and fixed variance parameters σ_c and σ_r, increasing the center size increases the absolute value of the *SNR*, which leads to better power. This also reflects the fact that there is more precision in

Table I. Setup for simulations in the balanced case.

Number of centers	$N_c = 200$
Number of atypical centers	$N_{atypical} = 1, 4, 20$
All center sizes	$N_i = N = 50$ (balanced)
Null model structure	$\mu = 0; \sigma_c = 1; \sigma_r = 4$
Alternative model	Shifted mean: $\mu + \delta$ (no changes in residual variance)
Number of replications	$N_{sim} = 2000, 500, 100$ ($N_{sim}N_{atypical} = 2000$)

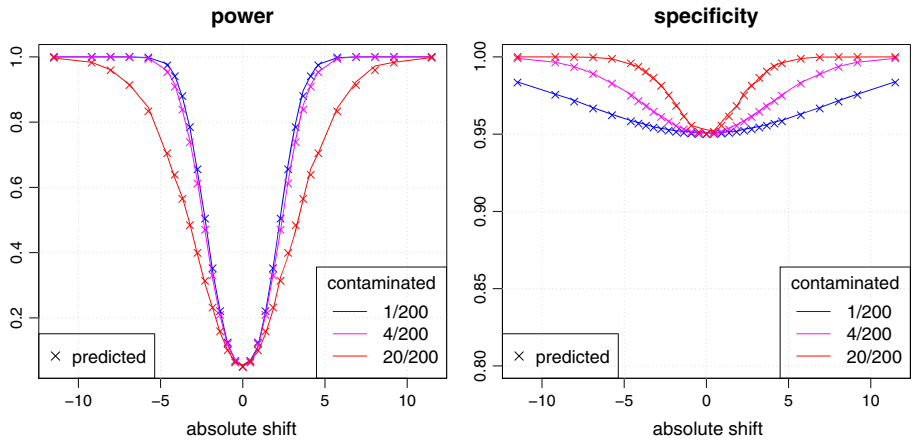

Figure 4. Average power (left panel) and average specificity (right panel) for the simulations described in Table I. The curves represent observed performance and the crosses represent predicted performance.

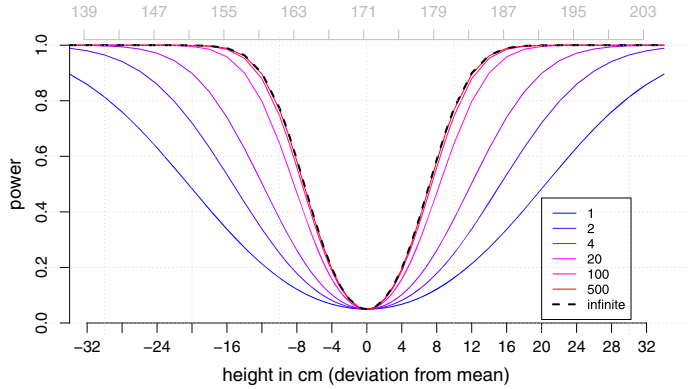

Figure 5. Theoretical power for the detection of shifts in height for 1% contamination and different atypical center sizes. Shifts are reported in the original scale of the model. Absolute height values are reported on a second x-axis on top of the plot. The dashed line represents the predicted power based on δ/σ_c as *SNR*.

the estimate of the location for a larger center. With small centers, an atypical location may occur more easily by the play of chance.

We illustrate this using the height (of the patient) in an actual clinical trial. Parameter estimates for the null model are $\mu = 171.35, \sigma_c = 3.55$, and $\sigma_r = 9.48$ (all expressed in cm), assuming all centers are clean and of equal size. We look at the predicted power if a signal were present in 1% of the centers, using expression (2). The power depends directly on the center size N in the *SNR* expression. The curves of power *versus* deviation of a center's location (from the mean μ) are plotted in Figure 5 for different values of the center size N. For arbitrarily large center sizes, the denominator of the *SNR* tends to σ_c, and the power reaches its maximum for all shifts. The power curve based on δ/σ_c can thus be seen as a benchmark curve describing the limiting performance for infinitely large sizes in the balanced case (thick dashed line in the figure).

4.3. Unbalanced setup and size effects

Because center sizes in actual clinical trials are rarely balanced, we now investigate through simulations how differences in center sizes affect the performance of the procedure. Instead of using fixed center sizes in the data generation step, we select random sizes drawn from a suitable discrete distribution. We use three distributions derived from observed sizes in three clinical datasets that are considered representative of many randomized trials. As shown on Figure 6, all three distributions are positively skewed: smaller centers are much more frequent than larger centers, with a few centers much larger than on average. There is however a clear difference between the three distributions in terms of the overall center size:

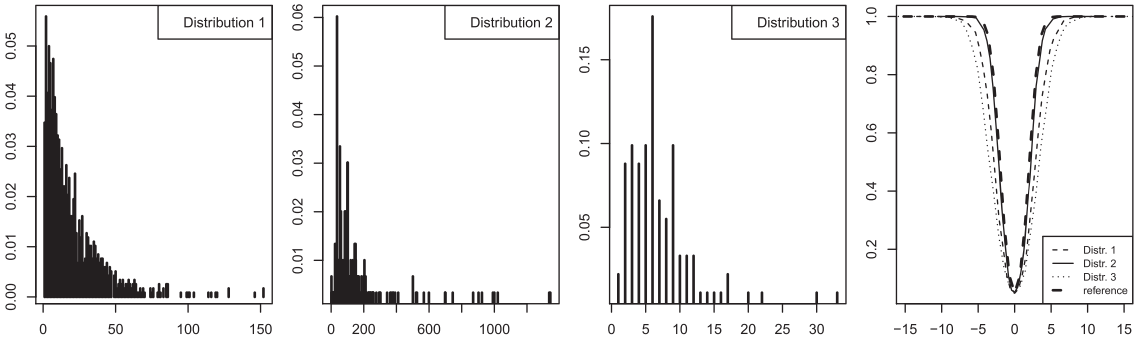

Figure 6. Three left panels: distribution of center size in three actual clinical trials. Right panel: average power in simulations with center sizes selected at random from these distributions.

Table II. Setup for simulations with center sizes selected at random.

Number of centers	$N_c = 100$
Number of atypical centers	$N_{atypical} = 1$
All center sizes	Random (from observed distributions in Figure 6)
Null model structure	$\mu = 0; \sigma_c = 1; \sigma_r = 4$
Alternative model	Shifted mean: $\mu + \delta$ (no changes in residual variance)
Number of replications	$N_{sim} = 2000$

Table III. Setup for simulations with varying atypical center sizes.

Number of centers	$N_c = 100$
Number of atypical centers	$N_{atypical} = 1$
Clean center sizes	Random (observed distributions in Figure 6)
Atypical center sizes	1%, 5%, 20%, 50%, 80%, maximum and 10 × maximum of the observed distributions
Null model structure	$\mu = 0; \sigma_c = 1; \sigma_r = 4$ (Figure 7) and $\sigma_r = 2$ (Figure 8)
Alternative model	Shifted mean: $\mu + \delta$ (no changes in residual variance)
Number of replications	$N_{sim} = 2000$

most centers were in the 1–100 range in the first trial, in the 1–500 range in the second trial, and in the 1–20 range in the third trial.

We carry out a second set of simulations with all center sizes, including one atypical center, taken at random from the distributions shown on Figure 6. Details of the setup are shown in Table II.

The average power as a function of the absolute shift is shown in the right panel of Figure 6. As could be expected, the power is the highest for the second distribution, and it is the lowest for the third. This effect of overall center size can be explained through the *SNR*. Unfortunately, unlike in the balanced case, it is not possible to define a unique *SNR* because we cannot capture the whole distribution of center sizes in one single effective size parameter. However, the *SNR* formula for the balanced case is still helpful to understand how larger overall sizes lead to higher power. In particular, when the order of magnitude of the sizes (say the median N_{med}) is such that σ_r^2/N_{med} is negligible against σ_c^2, the power curve will approximate the predicted power curve based on δ/σ_c (thick dashed line in the graph). This is the benchmark curve that we already saw in Figure 5.

In our last set of simulations, outlined in Table III, the size of the atypical center was held at meaningful levels with respect to the corresponding distribution of center size. For a small center size, we considered the 1%, 5%, and 20% quantiles of the distribution; for a medium size, we considered the median of the distribution; and for a large size, we considered the 80% quantile, the maximum of the distribution, and 10 times the maximum in order to assess the effect of an extremely large atypical center (beyond the range used for the clean centers).

The average power curves are displayed in Figure 7 (the color scale from blue to red indicates increasing size of the atypical center). The curves form a bundle, and the power, as expected, increases with increasing size of the atypical center. The overall size effect is also reflected in this plot as the position of

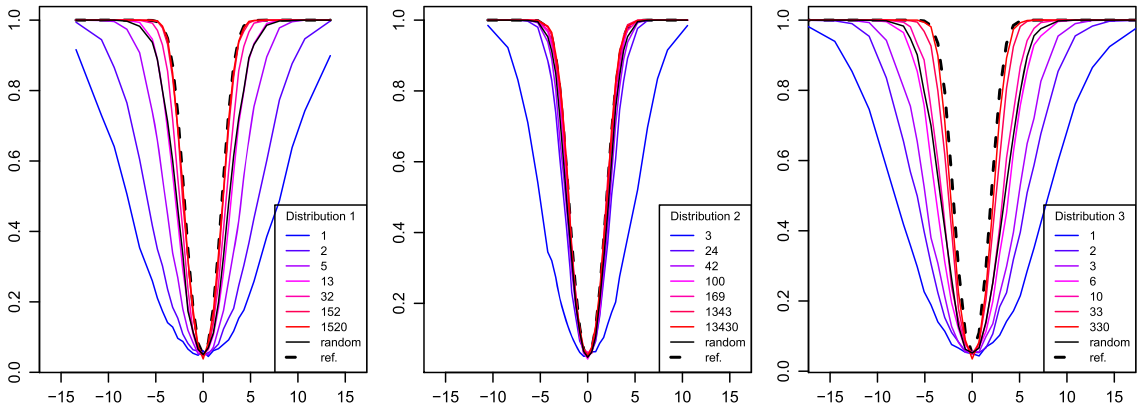

Figure 7. Power as a function of absolute shift with the first (left panel), second (middle panel), and third (right panel) size distributions shown in Figure 6. The color scale (blue to red) indicates the size of the atypical center (see legend). The black solid line corresponds to the case where all sizes are random, and the black dashed line is the benchmark line (prediction based on *SNR* of δ/σ_c).

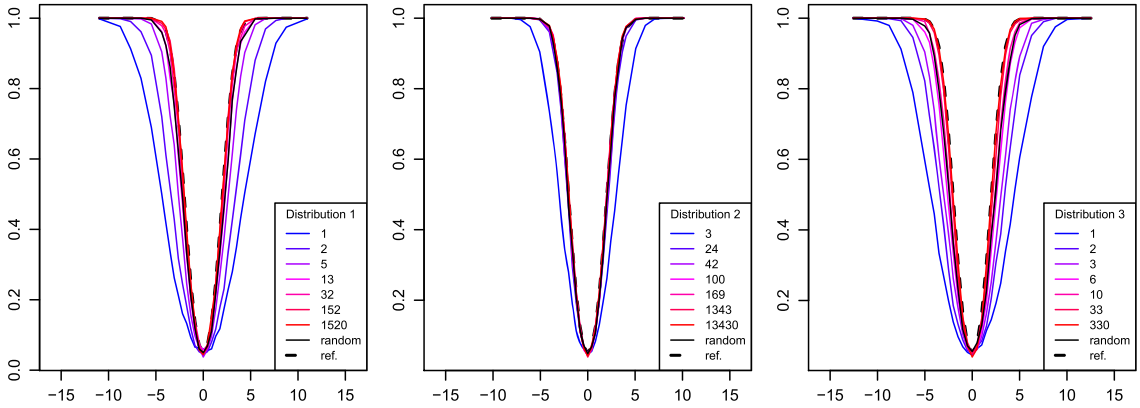

Figure 8. Same as Figure 7, but with $\sigma_c = 1$ and $\sigma_r = 2$.

the bundle depends on the overall center size (higher for the second distribution than for the third). The black line represents the power when all centers have random size (as in the right panel of Figure 6) and is almost superimposable with the curve for a median atypical center size.

We note that the effect of a relatively small atypical center (blue/bluish lines for atypical center smaller than median) is potentially more important than the effect of a relatively large one (red/reddish lines for atypical center larger than median). Indeed, because of the skewness of the size distribution, the lower quantiles correspond to rather small sizes, and these have a bigger impact on power. We also note the particular pattern in the middle panel of Figure 7: again, the power is the lowest when the size of the atypical center is smallest, but the lines for all larger sizes are quite close to each other. In this distribution, the overall size is rather large, and we rapidly approach the benchmark curve (based on δ/σ_c as *SNR*), which corresponds to the highest possible power (in the absence of size effects). With the third distribution (right panel of Figure 7), we do not approach the benchmark curve with the maximum of the distribution as atypical size, we only get close to it with 10 times this size.

Because the denominator $\sqrt{\sigma_c^2 + \frac{\sigma_r^2}{N_i}}$ is key in understanding the size effects, the values of σ_r and σ_c play an important role. However, the pattern seen in the graphs depends only on the ratio σ_r/σ_c. If this ratio gets smaller, the denominator decreases and the power increases for all center sizes. This is illustrated in Figure 8 where $\sigma_r/\sigma_c = 2$ rather than $\sigma_r/\sigma_c = 4$ as in Figure 7.

4.4. Practical interpretation of the simulation results

In this section, we focus on the prediction of the power of our detection procedure in practice. As shown in Section 4.2, this question can be addressed simply in a balanced setup, provided the parameters of the

Table IV. Overview of the effects of center size.

Balanced setup $N_{med} = N_i = N$	Exact prediction possible, using $SNR = \delta/\sqrt{\sigma_c^2 + \sigma_r^2/N}$
Unbalanced setup and $\sigma_r^2/N_{med} \ll \sigma_c^2$ (large overall size)	
$\sigma_r^2/N_i \ll \sigma_c^2$	Approximate prediction based on $SNR = \delta/\sigma_c$
$\sigma_r^2/N_i \not\ll \sigma_c^2$ (atypical size small)	Expect reduced power
Unbalanced setup and $\sigma_r^2/N_{med} \not\ll \sigma_c^2$ (small overall size)	
$\sigma_r^2/N_i \ll \sigma_c^2$ and small contamination rate	Approximate prediction based on $SNR = \delta/\sigma_c$
$\sigma_r^2/N_i \not\ll \sigma_c^2$ or larger contamination rate	Expect reduced power

null model are known. In this respect, it is important to emphasize that we need the parameters of the model for the clean data, because $\hat{\mu}$ and $\hat{\sigma}_c$ will likely be biased by the contamination (in particular, the center variance will tend to be inflated).

If the setup is not balanced, precise prediction is more difficult, but under certain conditions, we can still obtain reasonable approximations. An overview of the effects of center size is given in Table IV, where N_{med} denotes the median center size and N_i is representative of the size of the atypical center(s).

The size effect will be negligible when $\sigma_r^2 \ll \sigma_c^2$: the power curves for different atypical center sizes all coincide with the predicted power based on δ/σ_c. This is a rare case, but it might arise when there is a strong center effect, for example, in a multicenter study with important regional effects in the variable of interest.

If the center variance does not dominate the patient variance, size effects come into play: when N_i is small, we can expect low power, and when it is large, we can expect high power, but in both cases, the overall size of the clean centers plays a role (when the overall size is large, small atypical centers will be more easily detected, and when it is small, the power to detect large atypical centers may be inferior to the predicted power based on δ/σ_c).

Finally, it is important to realize that a power estimate is only an average detection rate (detection of a specific atypical center being a random event). In the important case in which exactly one center is atypical, the interpretation as a probability of detection is straightforward.

4.5. Departures from the assumptions

The location-shift model used in the former sections implies several assumptions: normality of the data, homoscedasticity, and a common systematic location shift in the atypical centers. While these are helpful in deriving the properties of the procedure, they are unlikely to hold in real trials. In this section we show that the procedure is reasonably robust with respect to departures from the two last assumptions.

This is investigated with a simulation study extending the setup in Table I and with a global contamination of 2% (4 out of 200 centers were atypical) where the shifts come in pairs of two. The case of opposite shifts ($\delta_1 = -\delta_2$) is compared with the standard case (identical shifts $\delta_1 = \delta_2$) in Table V, while asymmetric setups are considered in Table VI. In both tables, separate power evaluations $power_1$ and $power_2$ are given. In addition, a heteroscedastic scenario was implemented with a random patient standard deviation according to a Gamma distribution (shape $k = 32$ and scale $\theta = 1/8$) and magnified values (inflation factors $inflate_1$ and $inflate_2$) in the atypical centers reflecting the situation that a shift in

Table V. Multiple signals: symmetric case.

Standard ($\delta_2 = \delta_1$)			Symmetric ($\delta_2 = -\delta_1$)			Symmetric and varying variance				
δ_1	$power_1$	$power_2$	δ_1	$power_1$	$power_2$	δ_1	$inflate_1$	$inflate_2$	$power_1$	$power_2$
1	0.14	0.13	1	0.14	0.13	1	1.2	1.2	0.14	0.15
3	0.69	0.70	3	0.72	0.70	3	1.5	1.5	0.51	0.53
6	1.00	1.00	6	1.00	1.00	6	2	2	0.87	0.87
10	1.00	1.00	10	1.00	1.00	10	3	3	0.97	0.96

Table VI. Multiple signals: asymmetric cases.

Asymmetric				Asymmetric and varying variance			
δ_1	δ_2	$power_1$	$power_2$	$inflate_1$	$inflate_2$	$power_1$	$power_2$
−1	3	0.12	0.73	1.2	1.5	0.12	0.54
−3	6	0.66	1.00	1.5	2	0.50	0.89
−1	6	0.10	1.00	1.2	2	0.11	0.89
−6	10	0.99	1.00	2	3	0.84	0.96
1	6	0.08	1.00	1.2	2	0.09	0.88
6	10	0.99	1.00	2	3	0.84	0.96

location may come with a larger variability (in this scenario, center sizes were drawn at random from distribution 1 in Figure 6).

The few scenarios presented here do not provide a full coverage of all possible departures, but they do give insight in the behavior of the procedure. For the same global contamination rate, the fact of having opposite rather than parallel signals of comparable magnitude does not lead to reduced power (thanks to the fact that the hybrid model yields a better estimate of the location parameter of the null); when shifts have substantially different magnitude, however, the stronger one is the more easily detected. The fact of having heteroscedasticity and inflated patient standard deviation in the atypical centers reduces power, but not dramatically.

With the aforementioned results in mind, we can easily understand what happens when there are multiple atypical centers and shifts in the centers that are not identical but show some spread around a common systematic component. When this spread is rather small (say the standard deviation is at most one-eighth of the systematic shift), the performance will not be affected much (this was confirmed in simulations not reported here). Finally, we remark that specificity remains conservatively controlled across all scenarios.

5. Application to actual clinical trials

We first illustrate the use of our procedure using data on vital signs in a multicenter international clinical trial that accrued patients in 218 centers. This is an on-going trial in which CSM revealed a problem with measurements of the body temperature of the patients in 10 centers (shown in Figure 9) from the same country. Upon closer inspection, it turned out that there was an issue with the calibration of a lot of thermometers used in that country. The miscalibration resulted in measurements that were systematically shifted down, though still being within the allowed range (and thus passing the range checks that are routinely applied to clinical data). On-site monitoring could easily have missed small shifts in body temperature at each patient visit, but the accumulation of low temperatures led to highly significant differences between the centers using miscalibrated thermometers and the others. Specifically, the procedure correctly detected 10 out of the 12 atypical centers (power of 83%) and falsely detected two out of 206 clean centers (specificity of 99%).

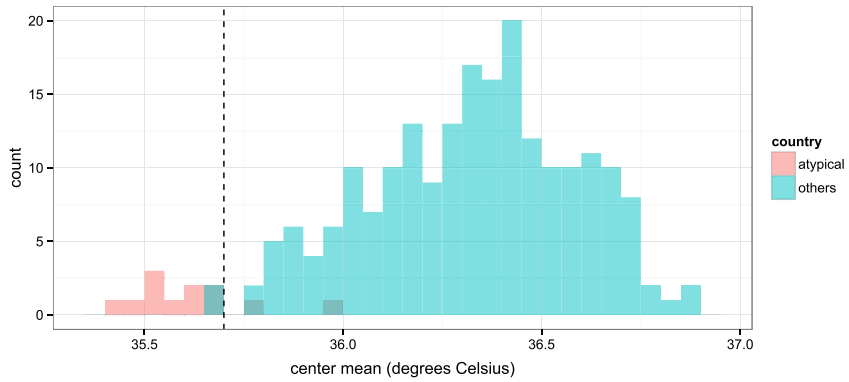

Figure 9. Histogram of mean temperatures (in degrees Celsius) in the 218 centers of a multinational clinical trial with two groups: the 12 centers from the country with the miscalibration issue and the 206 other centers. The vertical line delimits the region of detection.

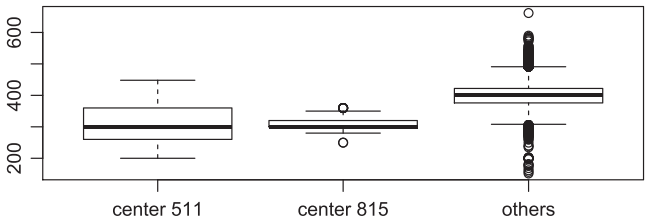

Figure 10. Boxplot of QT interval durations (in ms) in centers 511 and 815 compared with the other centers.

We take our second example from the field of cardiology. In this trial, 2364 patients were recruited in 235 centers. One of the variables of interest was the measurement of the QT interval, which is the duration of a clinically relevant portion of the heart's electrical cycle. The QT interval was measured across 11 visits. Two centers were identified as being atypical (center 511, $p = 1.4 \times 10^{-6}$, and center 815, $p = 5.5 \times 10^{-7}$). These centers, indeed, have atypical ranges for the QT interval duration, as shown in Figure 10, and the detection procedure correctly pinpointed centers where the measurements were clearly different from the others. This might trigger further investigations as to whether the differences are related to a different population, to a difference in the measurement process, to some other specific cause, or to the play of chance. In this case, the signal was attributed to population differences and was not considered a critical finding in itself. Nevertheless, in conjunction with signals obtained from other tests, it contributed to discovering the two centers as atypical. The principle of CSM is that the p-values from several tests are combined to compute a single overall score for each center.

We note in closing that in the two examples just discussed, the visit effect was considered a fixed effect in the mixed-effects model, which can easily be accommodated in the procedure.

6. Discussion

In this article, we have described a detection procedure that may be used as a component of CSM. This procedure detects whether a center is atypical, in terms of the mean of some continuous variable, by comparing it with all other centers. We have shown that the procedure responds to reasonably modest amounts of contamination, and we have shown that its performance depends on the SNR. We have also shown that the type I error probability is conservatively controlled and the minimum specificity can be set by means of the significance level α.

The advantage of the proposed procedure is that it does not require previous information about the parameters or the existence of atypical centers. We have shown through simulation that the performance is reasonably robust with respect to departures from the homoscedasticity and identical multiple shifts assumptions. In particular, shifts in opposite directions are detected with higher power than parallel signals of comparable magnitude. However, because the simulation conditions may not always be representative of scenarios that can be encountered in real cases, an additional validation via multiple real cases is needed to confirm the usefulness of the method.

The issue of non-normality is delicate as it may not always be possible to verify the assumptions for the location-shift model based on the aggregated hybrid data. If the location-shift normal model holds, one may exploit the fact that the datasets in individual centers are normal (albeit possibly shifted), or one may consider the centered datasets (by subtracting center means) together as an aggregated normal sample. However, because the log-normal model may be more appropriate than the normal model for certain biomedical variables, it may be worthwhile to perform a log transformation as a preprocessing step. We suggest that the transformation is carried out when there is evidence that the log-normal model is more plausible than the normal in the majority of the centers. If it turns out that neither model is likely, another transformation may be attempted, but one should be cautious in drawing conclusions in the presence of large deviations from normality.

The procedure extends to unbalanced setups, with the power for detection being lower when the atypical center is small. Because in small centers the estimate of the location is less precise, it is not possible to enhance the detection of atypical small centers without increasing the type I error probability. This implies that it may be more difficult to detect an atypical center at an early stage of enrollment and suggests that it may be worthwhile to run the procedure at different points in time, with increasing sample sizes.

Statistical detection procedures such as the one discussed in this paper are at the heart of CSM systems that combine the p-values obtained for many variables in order to detect atypical centers in the multidimensional space induced by these variables. The vectors of p-values obtained from several test procedures can be combined in an overall score for each center, in such a way that centers that were detected because of the play of chance are downplayed, while centers that were flagged multiple times across a number of tests are confirmed to be significantly different from all other centers [5]. The derivation of an overall score is beyond the scope of this paper, but the detection procedure presented here is of interest in so far as it does not require the distributional assumptions and/or the sample size requirements that are called for by many multivariate techniques. An ideal CSM system should be versatile and independent of the trial at hand and should not require prior information about any of the variables to be analyzed. This is in contrast to the setting of statistical quality control where information is available on the desired properties and specifications of a product to be controlled. In clinical trials, the situation is more complex as a result of the dimensionality (number of variables collected), diversity (type of variables collected), heterogeneity (systematic differences between centers and center sizes), and reasons why centers may be atypical, because discrepancies in the data may be as diverse as the reasons that caused them (transcription errors, measurement issues, misunderstandings, procedural problems, data tampering, or even data fabrication).

Appendix

We derive formulas (1) and (2). The hybrid sample can be considered as drawn at random from a mixture density whose probability density is a convex combination of the normal densities of the individual components:

$$f_{hybrid} = (1-w)f_0 + wf_1,$$

where f_0 is the density of the null distribution, f_1 is the density of the alternative distribution, and $w = n_1/(n_0+n_1)$ is the contamination rate. The parameters μ_{hybrid} and σ^2_{hybrid} of this density can be expressed in terms of the parameters of the components:

$$\mu_{hybrid} = (1-w)\mu_0 + w\mu_1, \text{ and } \sigma^2_{hybrid} = (1-w)\left\{(\mu_0-\mu_{hybrid})^2+\sigma^2\right\} + w\left\{(\mu_1-\mu_{hybrid})^2+\sigma^2\right\}$$

Consider now the case where $\mu_1 = \mu_0 + \delta$; then,

$$\mu_{hybrid} = \mu_0 + w\delta \quad \text{and} \quad \sigma^2_{hybrid} = \sigma^2 + \delta^2 w(1-w).$$

The probability of an observation being a false positive (an observation from the null distribution that falls in the tails of the hybrid distribution) can be computed as follows:

$$p_{FP} = P\left[\left\{X_0 < Q_{hybrid}(\alpha/2)\right\} \cup \left\{X_0 > Q_{hybrid}(1-\alpha/2)\right\}\right] \quad \text{with} \quad X_0 \sim N(\mu_0, \sigma^2),$$

where Q_{hybrid} refers to the quantile function for the hybrid model.

Similarly, the probability of an observation being a false negative (an observation from the alternative distribution that does not fall in the tails of the hybrid distribution) is then

$$p_{FN} = P\left[\left\{Q_{hybrid}(\alpha/2) < X_1 < Q_{hybrid}(1-\alpha/2)\right\}\right] \quad \text{with} \quad X_1 \sim N(\mu_0+\delta, \sigma^2).$$

Using the formula $Q_{hybrid}\left(\frac{\alpha}{2}\right) = \mu_0 + \sigma\left\{\frac{\delta}{\sigma}w + \sqrt{1+\left(\frac{\delta}{\sigma}\right)^2 w(1-w)}Q_Z\left(\frac{\alpha}{2}\right)\right\}$, where Q_Z is the quantile function for the standard normal distribution, leads to formulas (1) and (2).

Acknowledgements

The authors gratefully acknowledge financial support from the Walloon Government under the BioWin framework (consortium agreement no. 6741). The authors would also like to thank Bernadette Govaerts and Christian Ritter from Université Catholique de Louvain for insightful discussions and the reviewers for making suggestions that significantly improved the paper. This research (Catherine Legrand and Tomasz Burzykowski) is supported by the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy).

References

1. Eisenstein EL, Collins R, Cracknell BS, Podesta O, Reid ED, Sandercock P, Shakhov Y, Terrin ML, Sellers MA, Califf RM, Granger CB, Diaz R. Sensible approaches for reducing clinical trial costs. *Clinical Trials* 2008; **5**(1):75–84.
2. Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, Lesaffre E, Murray G, Edler L, Hutton J, Colton T, Lachenbruch P, Verma BL. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in Medicine* 1999; **18**:3435–3451.
3. Guidance for industry: oversight of clinical investigations - a risk-based approach to monitoring, U.S. Department of Health and Human Services, Food and Drug Administration: Silver Spring, MD, August 2014.
4. Reflection paper on risk based quality management in clinical trials. EMA/INS/GCP/394194/2011, European Medicines Agency: London, UK, August 2011.
5. 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. *Clinical Trials* 2012; **9**(6):705–713.
6. Gutttman I. Care and handling of univariate or multivariate outliners in detecting spuriosity: a Bayesian approach. *Technometrics* 1973; **15**(4):723–738.
7. Everitt BS, Hand DJ. *Finite Mixture Distributions*. Chapman and Hall: London, 1981.
8. Früwirth-Schnatter S. *Finite Mixture and Markov Switching Models*. Springer: New York, 2006.
9. Brown H, Prescott R. *Applied Mixed Models in Medicine*. John Wiley: Chichester, 1999.
10. Ohlssen DI, Sharples LD, Spiegelhalter DJ. Flexible random-effects models using Bayesian semi–parametric models: applications to institutional comparisons. *Statistics in Medicine* 2007; **26**:2088–2112.