Detection of Outlying Correlation Coefficients in Multicenter Clinical Trials

Non Peer-reviewed author version

Desmet, Lieven; Venet, David; Trotta, Laura; BURZYKOWSKI, Tomasz & BUYSE, Marc (2025) Detection of Outlying Correlation Coefficients in Multicenter Clinical Trials. In: Pharmaceutical statistics, 24 (3) (Art N° e70013).

DOI: 10.1002/pst.70013
Handle: http://hdl.handle.net/1942/45871

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Detection of outlying correlation coefficients in multicenter clinical trials

Lieven Desmet*[1,2] | David Venet[1,3] | Laura Trotta[4] | Tomasz Burzykowski[5,6] | Marc Buyse[4,5,6]

[1]joint first authors
[2]Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[3]Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium
[4]CluePoints, Louvain-la-Neuve, Belgium
[5]International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium
[6]Data Science Institute, Hasselt University, Hasselt, Belgium

Correspondence
*Lieven Desmet, Email: lieven.desmet@uclouvain.be

Present Address
Lieven Desmet,
ISBA-LIDAM, UCLouvain,
Voie du Roman Pays 20,
1348 Louvain-la-Neuve

Summary

Central statistical monitoring aims at finding centers whose data distribution differs significantly from the other centers in multicentric clinical trials. Such differences may point to data quality issues due to negligence, misconduct or fraud. Data distributions can be compared across centers in many different ways, depending on the type of data (*e.g.* numerical or categorical), whether a univariate or a multivariate comparison is performed, and so on.

In that framework, we present two methods aimed at detecting centers with outlying bivariate Pearson correlation coefficient. One of the methods directly compares the correlations across centers. The other method conditions the test on one of the marginal standard deviations, which makes the test on correlation independent of the centers standard deviations. Both methods are shown to perform equally well on simulated data. They are also applied on real world data, where they identify centers with outlying correlations. The findings of the two tests are compared, showing that they concord for centers with average standard deviation, but differ for centers with extreme standard deviation.

While the focus here is on central statistical monitoring, the methods are general and can be used in other settings.

KEYWORDS:
Correlation, statistical monitoring, multicenter clinical trial

1 | INTRODUCTION

Misconduct and fraud are serious concerns in clinical research. Data fabrication and falsification have been reported as two main cases of misconduct and fraud in clinical trials [1,2]. If making up clinical data with plausible univariate properties is a challenge, it is well-known that the multivariate structure of the data is even harder to imitate [1]. In particular, it has been shown that high correlation coefficients can be considered as a leading sign of data fabrication [3].

In this paper, we focus on the problem of detecting centers with outlying correlation coefficients in a multicenter clinical trial. This problem is of interest in the larger context of central statistical monitoring [4,5,6] (CSM) of clinical trials, where centers are tested on a large number of variables of potentially different nature. The discovered inconsistencies may point to data quality issues, and so are worth investigating. These tests are based on statistical models dependent on the variable type, and commonly consider one variable at a time, *e.g.*, patient's weight measured at some visit. For such a numerical variable, outlying centers may for instance be identified in terms of location and variability. We have however found [7,8] that, because of natural variability

[0]**Abbreviations:** CSM, central statistical monitoring

between centers, statistical models should include a random center effect. As a result of a test, each center is assigned a p-value reflecting its consistency with the full set of centers.

The process of CSM thus entails running a large number of tests on all available variables, taking into acount their nature and taking care of the multiple testing problem by appropriately summarizing all p-values resulting from these tests[4,9,10,11]. However, individual univariate tests may miss some features of interest. For example, even if both univariate distributions of height and weight for a given center are reasonable, observing a negative or a too perfect correlation in that center may raise suspicion. Thus, from a data-fabrication point of view, tests that focus on the multivariate structure of the data are a welcome addition to a CSM toolkit.

There has been interest in literature in the general problem of comparing population correlation parameters or even correlation matrices (see, *e.g.*, Larntz and Perlman[12] and references therein), but this is based on a global null-hypothesis of equality of multiple population correlation matrices, and without making the link with statistical monitoring. Koziol et al.[13] introduced a graphical method based on the Larntz and Perlman test to reveal patterns in electroencephalographic responses. Taylor et al.[14] address fraud detection in questionnaires based on multivariate correlation, but by using graphical and permutation techniques. However, a directly actionable test for CSM turned out to be lacking after exploration of some naive proposals based on the Fisher result as well as an iterative approach based on the Larntz and Perlman procedure. Therefore, we develop two new tests specifically designed for CSM.

The aims of this paper include (*i*) description of suitable data models, both in absence and in presence of outlying centers, (*ii*) formulation of various statistical tests for detection of outlying centers, (*iii*) investigation of the performance of these tests in a simulation study, comparing them with existing approaches as benchmark, and, finally, (*iv*) illustration of the methods in a number of real data cases. In all cases tests are on a pair of continuous variables modeled with a bivariate normal distribution in each center.

The paper is structured as follows. Section 2 describes suitable statistical models and formulates the detection problem. In Section 3, we describe the developed detection methods. Section 4 presents results from a simulation study. In Section 5, we apply the developed methods in two different settings: datasets with deliberately fabricated centers, and real datasets encountered in central statistical monitoring practice (four datasets). Section 6 closes the paper with a discussion of different aspects of the described methods, including their application in CSM. The proposed tests are implemented in the publicly available R package `CSMtests`.

2 | DATA MODELS AND DETECTION PROBLEM

2.1 | Multicenter correlated data

The starting point for formalizing the problem is the description of the null-hypothesis case, *i.e.*, the neutral setting where the centers have a consistent and plausible correlation structure described by a particular statistical model.

We propose a hierarchical model that describes the data in each center and, additionally, the correlation structure across centers.

We consider N centers, each center c having n_c observations (x_i, y_i) generated according to a bivariate normal distribution $\mathcal{N}_2$ with center specific parameters (ρ_c to be specified further on),

$$\begin{pmatrix} X_{ic} \\ Y_{ic} \end{pmatrix} \middle| \rho_c \sim_{i.i.d.} \mathcal{N}_2 \left(\begin{pmatrix} \mu_{X,c} \\ \mu_{Y,c} \end{pmatrix}, \begin{bmatrix} \sigma^2_{X,c} & \rho_c \, \sigma_{X,c}\sigma_{Y,c} \\ \rho_c \, \sigma_{X,c}\sigma_{Y,c} & \sigma^2_{Y,c} \end{bmatrix} \right). \tag{1}$$

We use capital letters for the random variables and write (x_i, y_i) for the observed data. Recall the definition of the correlation coefficient,

$$r_c = \frac{\sum_{i=1}^{n_c}(x_{ic} - \bar{x}_c)(y_{ic} - \bar{y}_c)}{\sqrt{\sum_{i=1}^{n_c}(x_{ic} - \bar{x}_c)^2 \sum_{i=1}^{n_c}(y_{ic} - \bar{y}_c)^2}}.$$

From a well-known result due to Fisher[15,16], we can describe the sampling variability of r as follows. Across independent replications of the model (1) with (sufficiently large) size n, the Fisher z-transformed correlation

$$z(r_c) := \frac{1}{2} \ln \frac{1 + r_c}{1 - r_c} = \operatorname{atanh}(r_c) \tag{2}$$

approximately behaves as a normally-distributed random variable with mean $z(\rho_c)$, related to the ρ_c parameter in (1), and standard deviation $\frac{1}{\sqrt{n_c-3}}$, related to the sample size n_c. In short,

$$z(R_c) \mid z(\rho_c) \sim \mathcal{N}\Big(z(\rho_c), \tfrac{1}{n_c-3}\Big)\,. \tag{3}$$

In view of this result, n_c is supposed to be at least 5 to improve precision and avoid numerical problems.

To include a random effect, it is natural to use the transformed correlations. Therefore, we apply the random effect on the true (but unknown) center correlation parameters with a normal distribution, thereby allowing for random deviations to capture unobserved differences in center-specific populations, as in [7,8]. Formally the ρ_c parameters are drawn from the random variable P_c:

$$z(\mathrm{P}_c) := \frac{1}{2}\ln\frac{1+\mathrm{P}_c}{1-\mathrm{P}_c} \sim \mathcal{N}\Big(\mu_\rho, \sigma_\rho^2\Big)\,. \tag{4}$$

Using both (3) and (4), we obtain the complete distribution of the observed correlation coefficients:

$$z(R_c) \sim \mathcal{N}\Big(\mu_\rho, \sigma_\rho^2 + \tfrac{1}{n_c-3}\Big)\,. \tag{5}$$

2.2 | Detection problem

Our goal is to detect centers that are not compatible with the null model defined by (4) in the sense that they have an outlying correlation value (on either side, *i.e.*, too high or too low). One can imagine many scenarios of outlyingness, but we assume that N_0 centers (a majority) comply with (4), where $\mu_\rho := z(\rho_0)$ for some parameter value ρ_0, while N_1 centers (a minority) do not.

One possible scenario is that the latter obey an alternative normal model, with the same variance but a shifted mean that corresponds to $z(\rho_1)$. This will be a convenient setting for our simulation study, but by no means implies that our methods require that all alternative centers have the same mean correlation parameter. In what follows, the symbols ρ_0 and ρ_1 are used in the sense of the true parameters defining, respectively, the null model and an alternative model. An example of model densities for ρ_c is shown in Figure 1.

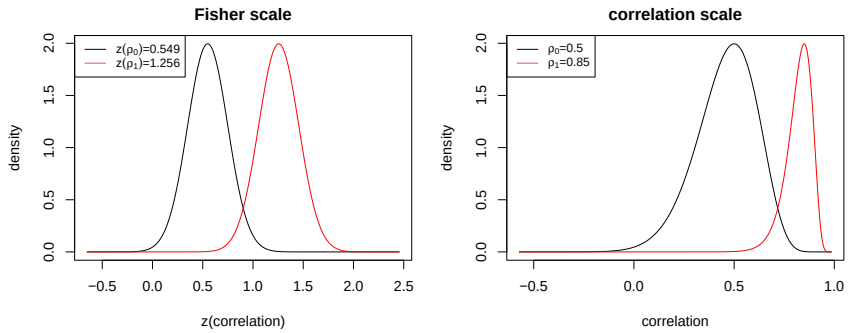

FIGURE 1 Left panel: Normal densities for the null and alternative models for the population correlation parameter on the Fisher scale (parameter $z(\rho_c)$). Mean parameters in this example correspond to ρ_c values of 0.5 for the null and 0.8 for the alternative, while the standard deviation is 0.2 in both cases. Right panel: on the direct correlation scale, distributions are deformed and asymmetric.

The purpose of a correlation test is to detect the N_1 outlying centers among all the $N = N_0 + N_1$ centers. This is done by assigning to each center a *p*-value that reflects its compatibility with the null model, and flag a center (*i.e.*, declare it as outlying) when this *p*-value is too small, which we define as below a threshold of 0.05.

If all centers conform to the null model, the *p*-values should have a uniform distribution, and flagging centers that have a *p*-value under the 0.05 threshold should lead to a procedure where on average 5% of the centers would be falsely rejected (false positives). When outlying centers are present, it is specifically those that should get significant *p*-values while the specificity should remain controlled.

3 | DETECTION METHODS

We first present a standard method based on the usual test to compare correlations, which does however not take into account the random effect. We then propose our two methods, one based on direct maximum likelihood estimation of the model, and another one that addresses the fact that sample variances and sample correlations are not independent. To conclude this section, we introduce an ad-hoc proposal based on the comparison of correlation matrices.

3.1 | Naive Fisher scale method

The commonly proposed test to compare correlations is to transform correlations to the Fisher scale, and then use (3) to derive the p-values, as implemented in the R package cocor[17]. We adapted this simple method to test each center individually in a 1-vs-all fashion. That is, we compared the correlation obtained only using data from that center r_c to the correlation obtained for all other centers r_{nc}. Using (3), under the hypothesis of no difference between centers

$$z(R_c) - z(R_{nc}) \sim \mathcal{N}\left(0, \frac{1}{n_c - 3} + \frac{1}{n_{nc} - 3}\right) \tag{6}$$

where n_c is the number of data points in the center, and n_{nc} the number of data points in other centers. We call this the "Naive Fisher" method.

3.2 | Fisher scale method

Our starting point is the distribution of the observed correlation coefficients (5), for which we consider the log-likelihood based on the observed data. Estimates $\hat{\mu}_\rho$ and $\hat{\sigma}_\rho$ are the arguments that maximize this log likelihood.

Each center gets its own score, given by $u_c = \frac{z(r_c) - \hat{\mu}_\rho}{\sqrt{\hat{\sigma}_\rho{}^2 + (n_c - 3)^{-1}}}$, and an associated p-value by using the standard normal reference and proceeding as in a two-sided hypothesis test, *i.e.*, we compute $p = 2\min\left(P(Z < u_c), P(Z > u_c)\right)$, where Z is a standard normal variable. With this p-value, the decision to flag a center is merely a question of setting a suitable threshold.

This approach is reasonable provided that that the model estimated from the data is a close enough approximation of the null-model, which is usually the case if the outlying centers only represent a small portion of the data (say, up to $5 - 10\%$). Alternatively, when many centers are outlying, this will tend to increase the estimated variance, leading to a conservative test.

3.3 | Fixed margin method

It is known that in the case of a sample drawn from a bivariate normal distribution, the distribution of its margin variances and the distribution of its correlation are not independent[18,19]: samples with larger variances on the margins have larger correlations, on average. This implies that a test of sample variances and a test of correlations would not be independent. This is further compounded by the observation that in real situations, individual centers often include patients from a subpopulation that has different characteristics than the overall population across centers[7], leading to large differences in variance across centers.

For instance, considering patient's height and weight, some centers could be located in cities that have a relatively homogeneous population (hence, little variability in height and weight, and a modest correlation between them), while other centers could have a more heterogeneous population (and a stronger correlation between height and weight). Since the test of correlation is meant to be used as one of many available tests, among which some specific tests of marginal variances, it may be warranted to construct a test of correlation that is independent of the marginal variance. The fixed margin approach takes into account the center-specific marginal variance by considering one of the margins (in our example, height or weight) as fixed in the bivariate normal model. In doing so, we need to adjust the distribution of the sample correlation coefficient. As developed in the Supplementary Information, a suitable transformation of the correlation coefficient follows a non-central t-distribution, with center dependent noncentrality parameter θ_c.

In the practical implementation of the fixed margin approach we introduce a random effect, *i.e.*, a normal model for the center specific ρ_c and proceed as follows. Parameters $\mu_\rho^*, \sigma_\rho^*$ are obtained by numerical optimisation, *i.e.*, as $\arg\max_{\mu_\rho, \sigma_\rho} \mathrm{L}$, where

$$\mathrm{L} = \sum_{c=1}^{N} \log \left[\int_{-\infty}^{\infty} \left\{ \varphi(z_c; \mu_\rho, \sigma_\rho) \psi_t(t_c; \tanh(z_c), x_{ci}, \sigma_{Xc}) \right\} dz_c \right] \tag{7}$$

with $z_c = z(\rho_c)$, $t_c = sign(r_{xY}) \sqrt{(n_c - 2) r_{xY}^2 / (1 - r_{xY}^2)}$ for center c, σ_{Xc} being the observed standard deviation for the center c, φ denoting the normal density, and ψ_t denoting the non-central t density (with parameter θ_c). The numerical integration on the full real axis is performed on a limited region determined by searching around ρ_c for a region of positive density, and extending that region until the density becomes very small. Although we do not prove that there is only a single maximum of (7), neither in our simulations nor with real data we were locked in a local optimum, as in that case we would have observed p-values that are globally far from uniformly distributed, leading to discrepancies in either sensitivity or specificity, which we did not observe. The p-value per center is similarly obtained by numerical integration:

$$p_c = \int_{-\infty}^{\infty} \left\{ \varphi(z_c; \mu_\rho, \sigma_\rho) \Psi_t(t_c; \tanh(z_c), x_{ci}, \sigma_{Xc}) \right\} dz_c, \tag{8}$$

where Ψ_t is the non-central t cumulative distribution function (with parameter θ_c).

The fixed margin approach, by construction, comes in two variants, depending on which of the variables plays the role of the independent (fixed) variable. We denote the associated p-values p_{xY} (fixing X) and p_{yX} (fixing Y), respectively. We can then define, in addition to the simple tests based on p_{xY} and p_{yX} directly, two additional combined tests, based on taking the minimum or the maximum of these p-values: $p_{\max} = \max(p_{xY}, p_{yX})$ and $p_{\min} = \min(p_{xY}, p_{yX})$. Obviously, in these combined tests, the former will be the more conservative and the latter the more liberal option.

3.4 | Iterative Larntz-Perlman test (for the purpose of comparison)

The method is based on the test of the equality of correlation matrices proposed by Larntz and Perlman [12], known for its advantageous small-sample properties. In the k-sample bivariate case, the null hypothesis is the equality of the k population correlation coefficients. The test statistic is based on the z-transformed empirical correlation coefficients. We turn this into a test of individual centers as follows. If the null is not rejected, no centers are flagged. If the null is rejected, we look for the center that contributes most to the deviation in the test statistic, flag it, and remove it from the data set. We run the test again, proceeding as above until the null is no longer rejected. At the end of the procedure we will have obtained a subset of flagged centers, if any. While we do not obtain p-values, we will be able to compare the test with our own proposals and compute performance criteria, based on which centers are flagged. We refer to this procedure as the "Larntz-Perlman" test.

4 | SIMULATION STUDY

4.1 | Simulation setup

Simulations were carried out according to the simple example setup introduced in Section 2. We generated data for $N = N_0 + N_1$ centers, where N_0 centers followed model (4) with parameter ρ_0 (null correlation), while N_1 centers followed the same model with a different parameter ρ_1 (alternative correlation). The contamination rate was $\phi = N_1/(N_0 + N_1)$. Data were generated according to model (1) with $\mu_{X,c} = \mu_{Y,c} = 0$, $\sigma_{X,c} = \sigma_{Y,c} = 1$.

The number of centers was kept fixed at $N = 100$. Simulation scenarios were constructed by combining different values for the following parameters: ρ_0, ρ_1, σ_ρ, and ϕ as shown in the table of Figure 2. To allow analysis of the potential effect of center size, we randomly drew the center sizes n_c from one among three discrete distributions. Those distributions correspond to sizes observed in three real clinical datasets, but with centers of size strictly smaller than 5 removed, and are shown in the lower panels of Figure 2. It is understood that this setup is a limited setting in terms of scenarios and data model. This exercise is meant to gather first results and highlight important trends, and is not an exhaustive simulation study.

4.2 | Method evaluation

All methods assign a p-value to each center. Centers with a p-value smaller than the 0.05 threshold were flagged as outlying. Performance was measured in terms of sensitivity and specificity across a number n_{sim} of replications of a cycle in which we

Symbol	Factor	Number of levels	Set of values
ρ_0	Null correlation	11	0, 0.1, 0.2, ..., 0.8, 0.9, 0.99
ρ_1	Alternative correlation	21	−0.99, −0.9, ..., −0.1, 0, 0.1, ..., 0.9, 0.99
σ_ρ	Fisher scale s.d.	3	0.02, 0.2 and 0.5
ϕ	Contamination rate	6	1%, 2%, 5%, 10%, 20%, 40%

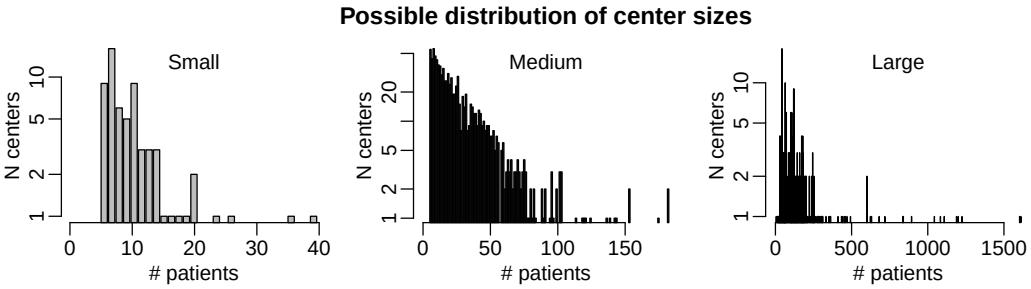

FIGURE 2 Setup for the simulation study. Top panel: table with the considered parameter values. Bottom panel: frequency plots for the sizes observed in three real cases, to be used as sampling probabilities when generating simulated centers. Small centers of size $N < 5$ were excluded from the simulation.

successively generate a multicenter data set, run the models, and obtain p-values allowing to make a decision per center. Across replications we gathered the total numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Sensitivity was computed as $TP/(TP + FN)$ and specificity as $TN/(FP + TN)$. We adapted the number of replications between 100 and 1000 to obtain a sufficient precision in the sensitivity estimate across the different values for the contamination rate.

The vertical grey lines are at $\rho_1 = \rho_0$, the horizontal grey lines indicate the 5% sensitivity that would be obtained by a random selection of centers (false positives). When assessing the performance and validity of different detection procedures, we examined a number of desirable properties: (i) sensitivity should be positively correlated with the absolute difference between ρ_1 and ρ_0 (for a given σ_ρ and center size distribution), and (ii) specificity should be superior to 95% across all scenarios (cf. the 5% threshold). Additionally, (iii) sensitivity should be high for small contamination rates and decrease with increasing rate (at ϕ =40%, say, one might argue that this can no longer be considered contamination). Taking into account the unbalanced nature of clinical trials, in addition, (iv) the specificity of the correlation test should not be influenced by the center size.

5% sensitivity that would be obtained by a random selection of centers (false positives).

4.3 | Selected results and discussion

Figure 3 presents performance curves for our two methods of interest and the naive Fisher method, as a function of the alternative correlation ρ_1 for $\rho_0 = 0.5$, for a contamination rate of 2%, which is the situation of greatest interest in practice, and different scenarios for the center sizes and the random effect standard deviation. These figures confirm that properties (i) and (ii) were met for both tests. There is little difference in terms of sensitivity between our methods, and both methods have specificity higher than the requested 95% level, as seen in Figure **??**. This is unlike the naive Fisher method, which has the highest sensitivity curves but a specificity lower than 95%, in particular for the larger center sizes or σ_ρ. In addition, we can assess the effect of center sizes (larger is better) and the random effect standard deviation (smaller is better).

Figure 4 (for sensitivity) and Figure 5 (for specificity) depict the combined effect of contamination rate, sizes distribution, and random effect size (σ_ρ) in 3 different (ρ_0, ρ_1) hypotheses for the naive method, the Fisher scale method and the fixed margin methods (both fixing x and its maximum variant).

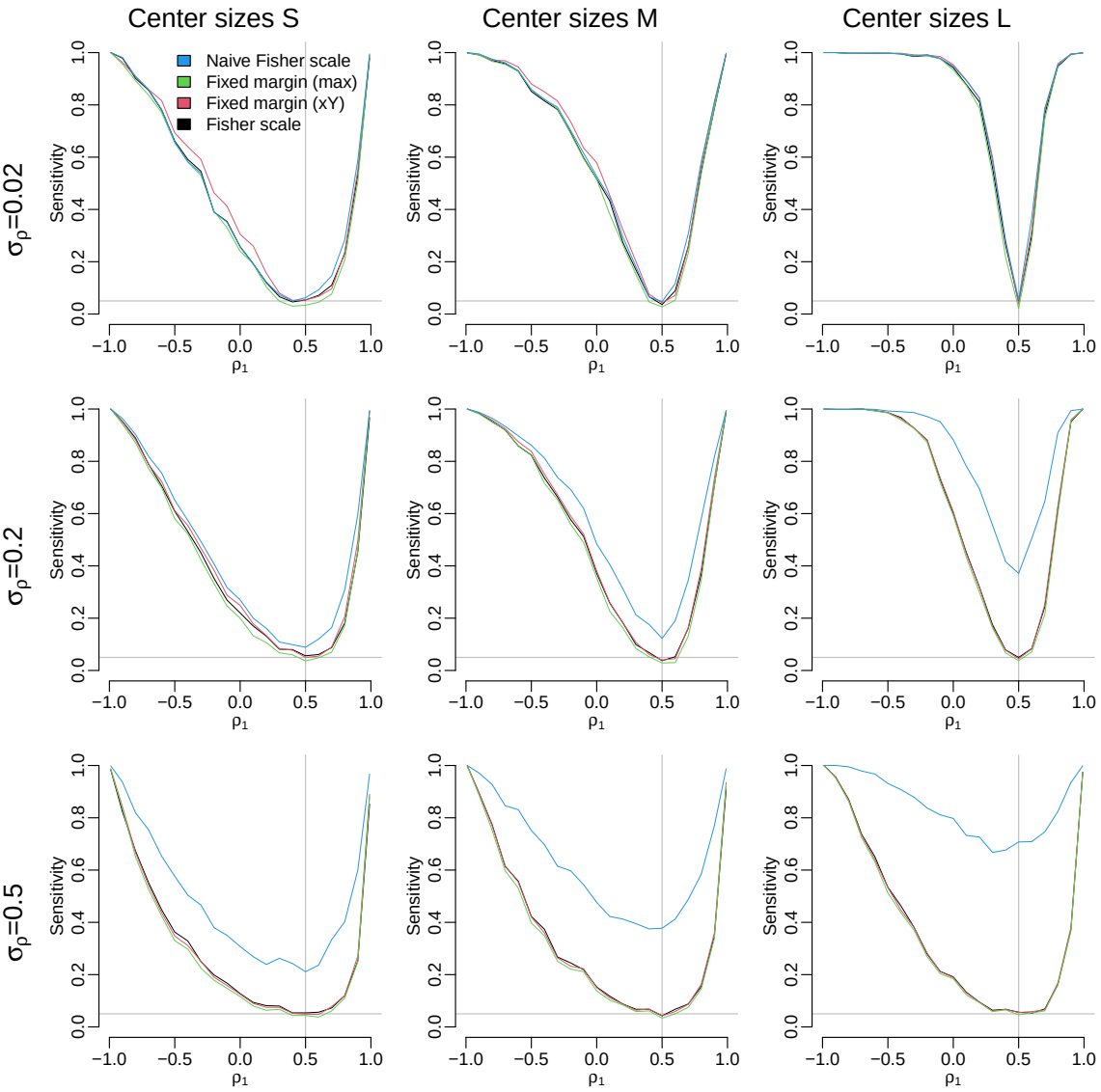

FIGURE 3 Curves of sensitivity versus alternative location parameter ρ_1 across different center size distributions (columns) and random effect sizes σ_ρ (rows), using $\Phi = 2\%$ contamination and $n_{sim} = 400$. Sensitivity increases for all methods, eventually reaching 1, as the alternative location moves away from the null position $\rho_0 = 0.5$, shown as a vertical grey line. At $\rho_1 = \rho_0 = 0.5$, the sensitivity for all methods but the naive Fisher scale corresponds to the 5% sensitivity that would be obtained by a random selection of centers (false positives), indicated by an horizontal grey line. A contrario, the naive Fisher Scale method has higher sensitivity with increasing σ_ρ, including when $\rho_1 = \rho_0$, where there is no difference between null and alternative centers.

First, the naive Fisher method is found to lack specificity (Figure 5) in the presence of medium or larger random effects, large centers, high contamination rate, or a combination thereof. This makes this test anti-conservative in many realistic settings. By contrast, the test has a very good sensitivity, which is logical as it basically flags everything.

Regarding the sensitivity of the Fisher and fixed margin methods, it is useful to make a distinction between two contamination ranges: low contamination (below 10%) versus larger contamination (above 10%). Focusing on the low contamination range, the sample size effect was quite strong, for a given σ_ρ, as the sensitivity increased markedly with sample size. For a given sample size, increasing σ_ρ decreased overall sensitivity. Remarkably, the reported methods behaved similarly and achieved a good level of sensitivity at these small contamination rates, with overall performance decreasing slightly with an increasing contamination rate. In the higher contamination range, sensitivity decreased more sharply, but not evenly across methods. For the Fisher scale

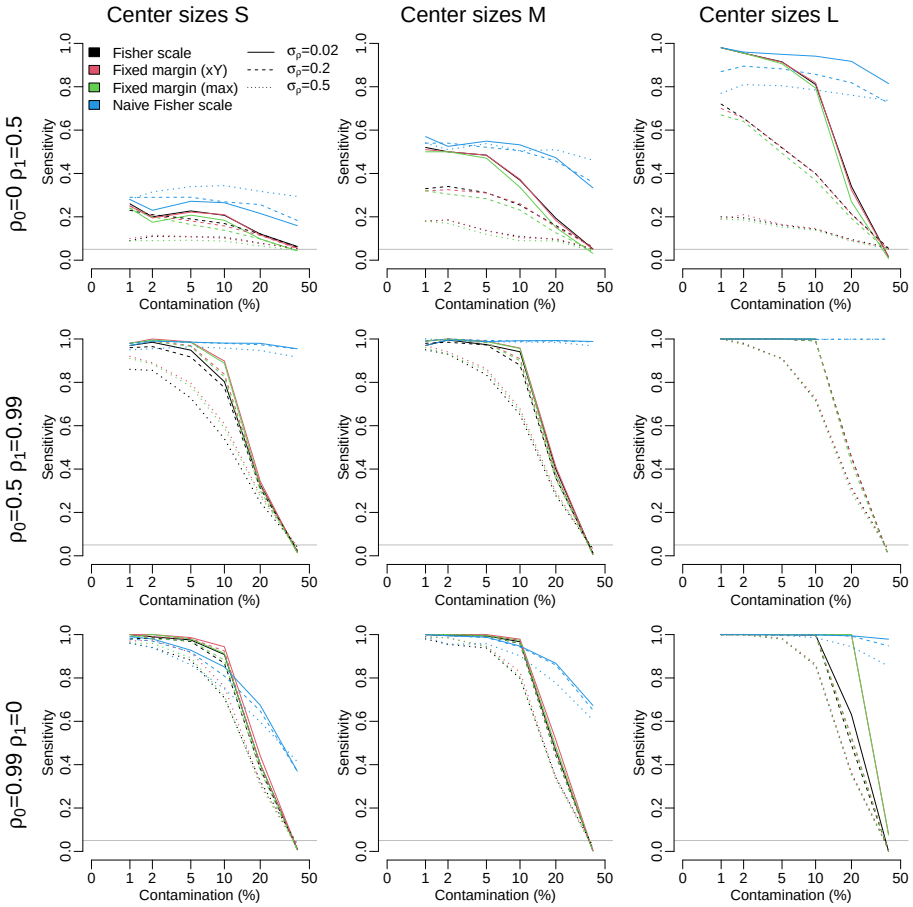

FIGURE 4 Sensitivity versus contamination rate for three illustrative values of ρ_0 and ρ_1 (rows), across three different center size distributions (columns). Several detection methods (Section 3) and three values of the random effect size σ_ρ (0.02, 0.2 and 0.5) are investigated (legend for colors and line types in top left panel). Sensitivity decreases as the contamination increases, reaching eventually the 5% value that would be obtained by a random selection of centers (horizontal grey line), except for the naive Fisher scale method that is more liberal. Sensitivity is globally higher for larger center size distribution, smaller random effect size and larger absolute difference between Fisher-transformed null and alternative correlations $|z(\rho_0) - z(\rho_1)|$.

method it decreased rather uniformly towards 0, whereas for the fixed margin method there remained substantial sensitivity in some scenarios, even at 40% contamination rate (mostly in cases where either ρ_0 or ρ_1 equals 0.99 or -0.99).

In terms of the specificity, these two methods were conservative, in the sense that their specificity remained consistently above the 95% threshold. In Figure 5, we can observe increasing specificity with increasing contamination rate.

Regarding the proposal based on the test of Larntz and Perlman, it lacks specificity (Figure **??**), especially with larger center sizes, even if it performs well in terms of sensitivity (Figure **??**), though not in line with property (*iii*). In fact, its response to an increasing contamination rate is opposite to what we expect for CSM applications and see for the two main methods above.

We then assessed the influence of center sizes on the tests, for sensitivity and specificity (Figures 6 and **??**). Specificity decreased with increasing center sizes for the naive Fisher and iterative Larntz-Perlman methods (the latter not reported in this text), while it did not vary with center sizes for the other methods. Sensitivity increased with center sizes for all methods.

In view of its robustness, the test based on p_{max} turned out to be the best option among the fixed margin test variants, as taking the maximum implies only a negligible loss of sensitivity with respect to the tests based on p_{xY} or p_{yX}.

In conclusion, in the reported simulations, both the Fisher scale and fixed margin methods exhibited good sensitivity, while they conservatively controlled the type I error probability.

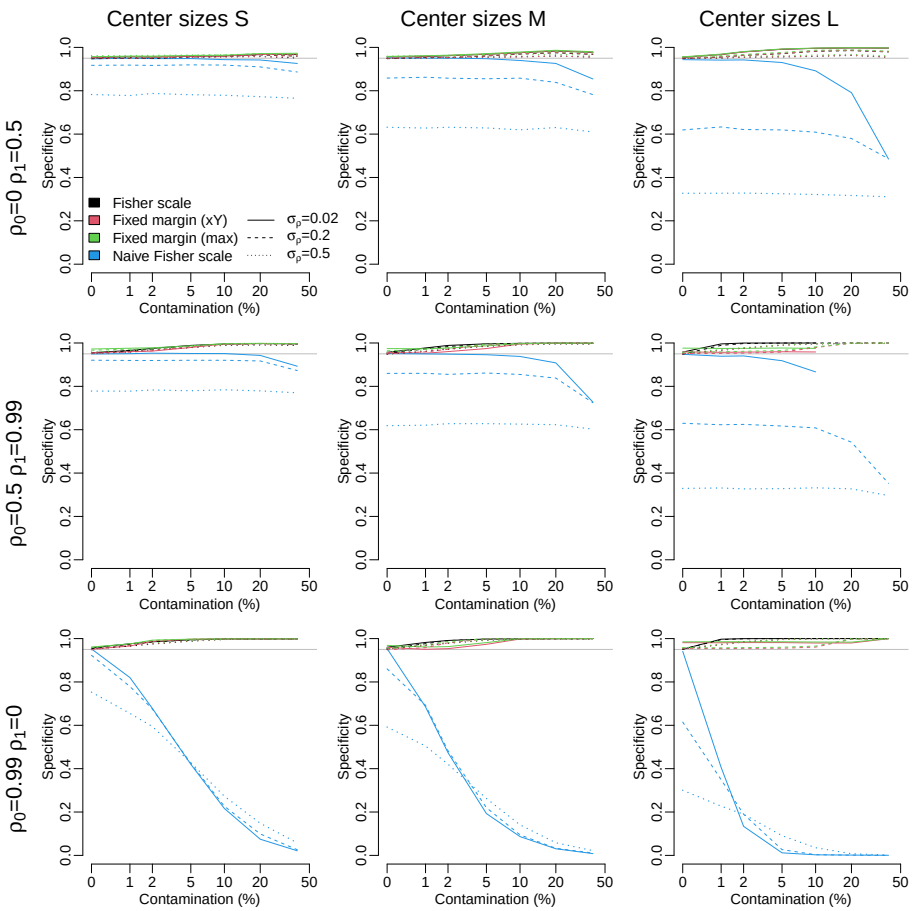

FIGURE 5 Specificity versus contamination rate for three illustrative values of ρ_0 and ρ_1 (rows), across three different center size distributions (columns). Several detection methods (Section 3) and three values of the random effect size σ_ρ (0.02, 0.2 and 0.5) are investigated (legend for colors and line types in top left panel). All methods are at least as conservative as the expected 95% specificity (horizontal grey line) across all scenarios, except for the naive Fisher scale method. The naive Fisher scale method becomes more liberal as the random effect size σ_ρ and/or the contamination increase.

4.4 | Differences between the Fisher and fixed margin methods

Even though both the Fisher and fixed margin methods performed similarly in terms of specificity and sensitivity, they are based on different assumptions and are bound to behave differently. The fixed margin methods are meant to decouple the correlation from the standard deviation. As per Hogben's result [19], higher correlations are expected for centers with larger marginal standard deviations, and lower correlations for centers with a lower marginal standard deviations. Since the Fisher method is directly based on the correlation, this implies that we would expect a correlation between *p-values derived from a 1-sided Fisher method test* and *the center standard deviations*. Such correlation should not appear for the fixed margin test.

To study this systematically, we generated datasets of 1000 centers of size 30 from a bivariate normal distribution, with correlation parameter (ρ_0) varying between 0 and 0.9, and various random effect standard deviations (σ_ρ). In Figure 7A can be seen that, as expected from Hogben's result, there was a substantial correlation between standard deviation and 1-sided Fisher test p-value when the correlation between the features (ρ_0) was high. This correlation between the observed standard deviation and p-values decreased with lower correlation ρ_0, as well as with larger random effect sizes (σ_ρ). However, when using the fixed margin test there was no such correlation between p-values and observed standard deviation (Figure 7B). This shows that the fixed margin test indeed decouples the test of the correlation from the effect of differences in variance between centers.

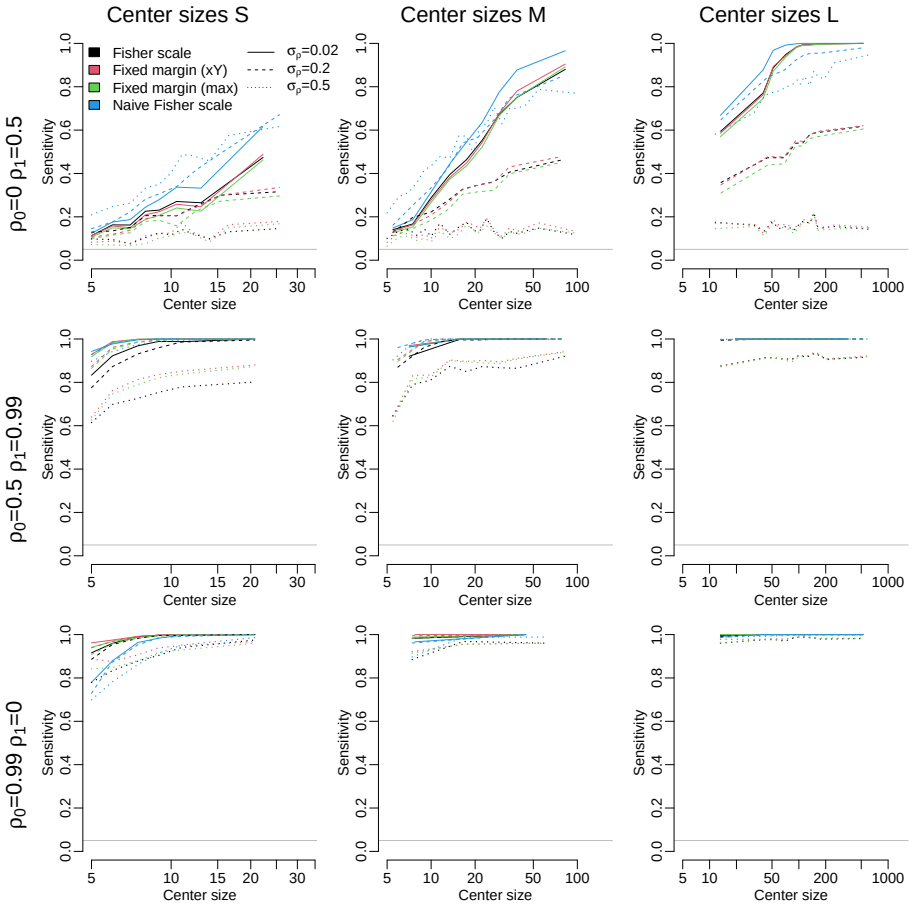

FIGURE 6 Sensitivity versus center size for three illustrative values of ρ_0 and ρ_1 (rows), across three different center size distributions (columns). Several detection methods (Section 3) and three values of the random effect size σ_ρ (0.02, 0.2 and 0.5) are investigated (legend for colors and line types in top left panel). Grey horizontal line indicates the baseline 5% sensitivity. Sensitivity increases with center sizes, the effect being larger for smaller σ_ρ, as for large σ_ρ most of the variability between centers is due to the random effect.

5 | APPLICATIONS AND CLINICAL EXAMPLES

In this section we apply the proposed tests to real-life datasets. The Fisher scale test and the maximum variant of the fixed margin test now become our default tests. We will refer to them as the *Fisher test* and *fixed margin test*, respectively, and we refer to their p-values with p_F and p_H (in honour of Fisher and Hogben), respectively.

5.1 | Detection of fabricated datasets

Akhtar-Danesh and Dehghan-Kooshkghazi[3] investigated how the correlation structure differs between real and fabricated clinical datasets. They considered two real bivariate datasets, for which the characteristics are given in Table 1, and for each one they had 34 medical faculty members make up a dataset of 40 fictitious patients mimicking the original correlation structure as closely as possible. It turned out that the investigators, in spite of detailed instructions and knowledge of the original data, failed to reproduce the data structure. Of interest to our research was whether these made up datasets could be identified as fabricated by a correlation test.

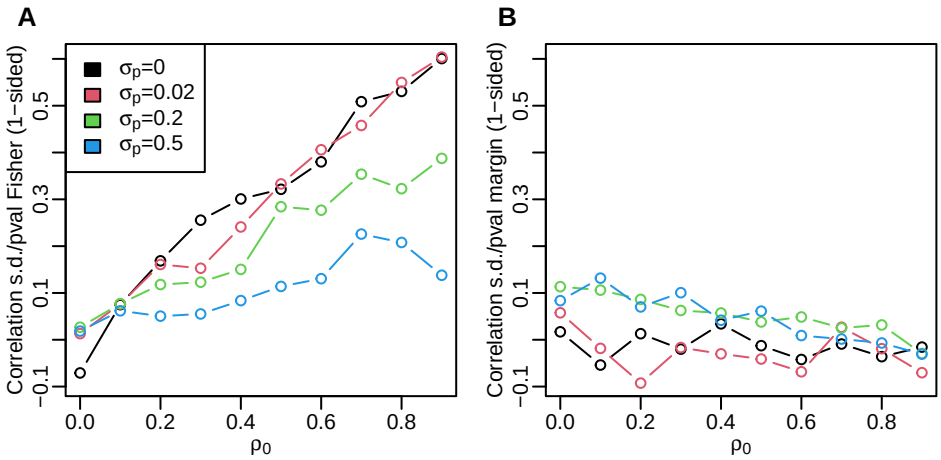

FIGURE 7 Spearman correlation between the p-value of 1-sided correlation tests and the standard deviation of a margin as a function of the null model correlation ρ_0 for (**A**) the Fisher test and (**B**) the fixed margin test. Curves for four random effect sizes σ_ρ are shown. For the Fisher test, the correlation increases with ρ_0, indicating that for larger ρ_0 centers flagged also have atypical marginal standard deviation, the effect being stronger for smaller σ_ρ. The fixed margin test does not show such a correlation, demonstrating that this test decouples correlation and standard deviation.

TABLE 1 Characteristics of the two original datasets that faculty members used to generate fake datasets.

Sample (size)	Variable X	Variable Y	Correlation
Students (65)	Height (in cm, $\bar{x}$=159.5, s_X=7.2)	Weight (in kg, $\bar{y}$=54.5, s_Y=9.2)	r=0.43
Newborn boys (637)	Gestational Age (in weeks, $\bar{x}$=40.1, s_X=1.0)	Birth Weight (in g, $\bar{y}$=3277, s_Y=443)	r=0.031

We focus on the first case (height-weight case) in Table 1, and designed a numerical experiment. In particular, we constructed hybrid multicenter datasets, in which 19 centers were simulated according to the bivariate normal distribution with true parameters corresponding to the values reported in the table (to be considered as the null model) and one center was taken from the list of fabricated centers (kindly supplied by the authors of [3]). For each of the 34 fabricated datasets we set up a simulation exercise by repeating the above 100 times. Based on a detection threshold of 0.05, sensitivity to detect the fabricated center and specificity were computed per fabricated dataset and reported in Supplementary Figure **??** (averages across 100 replications).

We concluded that the different correlation tests had comparable sensitivity. They also detected the majority of the fabricated centers (sensitivity was above 90% for half of the 34 fabricated datasets, and above 50% for two thirds of them). The centers with a correlation closest to that of the null model were, as expected, the most difficult to detect. We revisit this example in the Discussion to illustrate the idea of CSM.

5.2 | Clinical examples

We analysed data from existing clinical trials in different therapeutic fields. Clinical datasets typically have a large number of continuous variables, hence a huge number of pairs of variables to be tested, in multiple centers of different sizes. In addition, contrary to simulated data, real clinical data cannot be assumed to strictly follow an underlying bivariate normal model. In particular, there may be some extent of disparity in the center means or scales (*i.e.*, standard deviations of the X or Y variable).

In this analysis, we considered two pairs of continuous variables that are known to be positively correlated: diastolic and systolic blood pressure, and the height-weight pair. These variables are available in most clinical trials as part of the vital signs data collected for each patient.

The main questions we tried to answer are: *which centers are flagged* and *can this be interpreted in terms of the correlation structure*, and ideally can there be any plausible explanation as to *why centers are different*? We also investigated whether the different tests for outlying correlation yielded consistent results.

Four datasets from real clinical trials are presented. Their characteristics are given in Table 2 (only centers with at least 5 subjects and non-zero variance in both variables were included, after discarding incomplete cases). Trials were anonymized and were called A, B, and C. There were no suspicions of fraud or other serious problems in any of those trials. The correlations between features were moderate for the blood pressure pairs, and low for the height/weight pair. The observed random effects were moderate to weak for the height-weight pair but relatively high for the diastolic-systolic pairs. To fix ideas, for trial A, a center of infinite size would need to have a correlation above 75% to be considered as having a correlation significantly higher than the 57% mean correlation, while for the height/weight pair a correlation of 36% would be sufficient to become declared significantly above the 25% mean correlation.

TABLE 2 Details on pairs of variables from real trials on which correlation tests were applied. Random effect sizes were estimated using the Fisher test.

Trial name	Pair	#Patients	#Centers	Sizes range (median)	mean correlation	random effect (sd)
A	diastolic-systolic	4659	43	32-226 (100)	0.5688	0.17
B	diastolic-systolic	2711	240	5-140 (8)	0.5646	0.24
B	height-weight	2785	245	5-146 (8)	0.2554	0.059
C	diastolic-systolic	632	59	5-32 (9)	0.6759	0.11

As expected, center size played a major role, due to its influence on the sampling distribution of the correlation. Figure 8 presents the relationship between correlation and center size across centers for all four datasets. We can see that the limit of significance has a funnel-like shape, in particular for the Fisher method for which the 5% significance boundary lines are drawn. This is due to the variance used for the test: for small centers the main contributor to the variance is the sampling variance of the correlation, while for large centers the main contributor is the random effect. The funnel for the naive version is narrower, in particular for larger centers, leading to more centers being flagged, many being larger centers. At the limit of centers of infinite sizes, the naive method funnel would be a single point (the mean correlation), and would flag any center that does not have exactly that correlation, while for the Fisher method it would have a width proportional to the random effect standard deviation. A similar limit shape for the detection can be discerned for the fixed margin method, although it is not as clear cut because it also depends on the marginal variance. The Fisher and fixed margin tests flagged often the same centers (16 in common), while 17 centers were flagged by the Fisher method only and 4 centers were flagged by the fixed margin method only. The fixed margin method was more conservative as it used the highest p-value between the tests on both margins. The p-values given by both methods were correlated (Pearson correlations of the logarithm of the p-values were equal to 0.79, 0.83, 0.93 and 0.78 respectively for the 4 pairs of variables), showing again that the two tests gave similar results but with some discrepancies. In general, the Fisher method flagged a bit more than the expected 5% of the centers (7%, 5.4%, 4.9%, 8.5% respectively for the 4 datasets) while the fixed margin method was more conservative (2.3%, 3.3%, 2.9%, 6.8% respectively for the 4 datasets). This flagging rate was as expected for real-world studies that have no serious quality issues. The higher flagging rate for study C was likely due to a high level of discretisation in the values reported, which increases the sampling variance of the correlation, especially for small centers.

Figure 9 presents scatterplots of several particular cases for further detail. In trial A, a large center (A39) was detected with both methods with $p_H = 0.034$ and $p_F = 0.0059$ (size=226, correlation of 0.8211). This center has a higher correlation than expected and its marginal variances were not extreme, so it is flagged by both methods.

In trial B, a center (B225) of size 22 was detected for the blood pressure pair by the fixed margin test ($p_H = 0.025$), but not by the Fisher test ($p_F = 0.47$). As can be seen (Figure 9), this was because it had much smaller marginal standard deviations than the other centers (4.3 and 7.7 vs. median of 8 and 12.4, respectively), partially because most (13 out of 22) of its data points were replicated at (80,120). Because of this, a much lower correlation was expected by the fixed margin test, while this center still had a correlation that was higher than the mean. By contrast, using the height-weight pair in the same trial, a center (B143) of size 8 was detected by the Fisher test ($p_F = 0.039$) but not by the fixed margin test ($p_H = 0.16$). This was because it

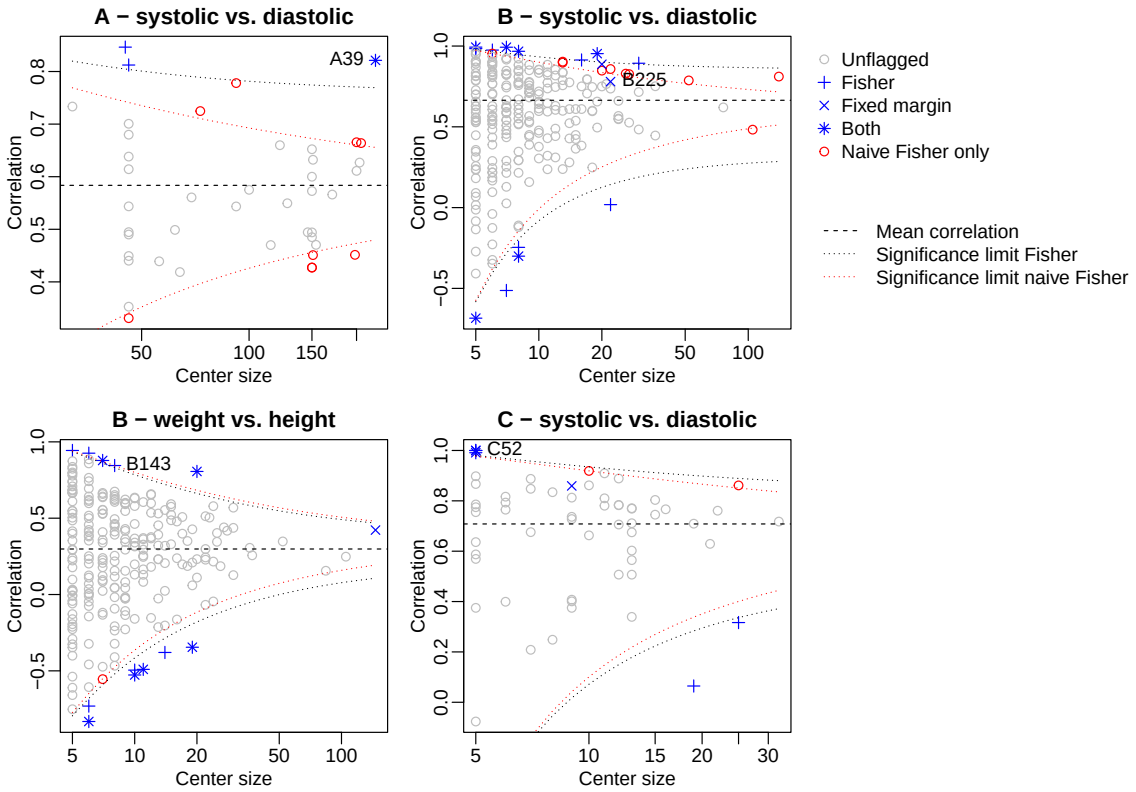

FIGURE 8 Observed correlations in larger centers tend to be more concentrated around the mean, as can be seen from the funnel-like patterns in the scatter plot of correlation versus center size in 4 example datasets. For infinite size centers, the funnels collapse to a single point for the naive Fisher scale test, while they have a certain range for the Fisher scale test. This makes large centers likely to be flagged by the naive Fisher scale test. There is a large overlap between the centers flagged by the Fisher scale and the fixed margin test, although some centers are flagged by only one method. There are no strong outliers in the data presented, as the significant centers are always relatively close to the limits of the funnel.

had much larger marginal standard deviations on both margins than the other centers (10.5 and 24.0 vs. median of 6.2 and 12.2, respectively). For this reason, the fixed margin test accounted for a high correlation already and this center was not given a low p-value, even though its correlation (at 0.84) is much higher than the average.

In center C52, 5 diastolic-systolic observations collapsed to 3 perfectly aligned points, resulting in extreme correlations and p-values, so that the center was flagged by both methods. This points to a general phenomenon: extreme correlations are sometimes obtained from a combination of small center sizes and overlapping datapoints due to discretisation (in this example: to multiples of 10 mmHg).

In summary, centers can be detected if their correlation is markedly outlying in either the positive or the negative direction with respect to the general tendency across centers and taking into account the center size. In most cases the two tests yielded similar conclusions, unless centers had extreme standard deviations on their margins.

6 | DISCUSSION

We proposed two different methods for the test of the correlation, with similar performance in terms of sensitivity and specificity in our examples and simulations, but based on different assumptions. The difference is that the fixed margin method conditions the test on the standard deviation of one of the features. This can be useful when the standard deviation varies a lot between centers, so that differences in standard deviation explain a large fraction of the differences in correlations. This is however only important if the average correlation is at least moderate and those differences are not negligible compared to random variation

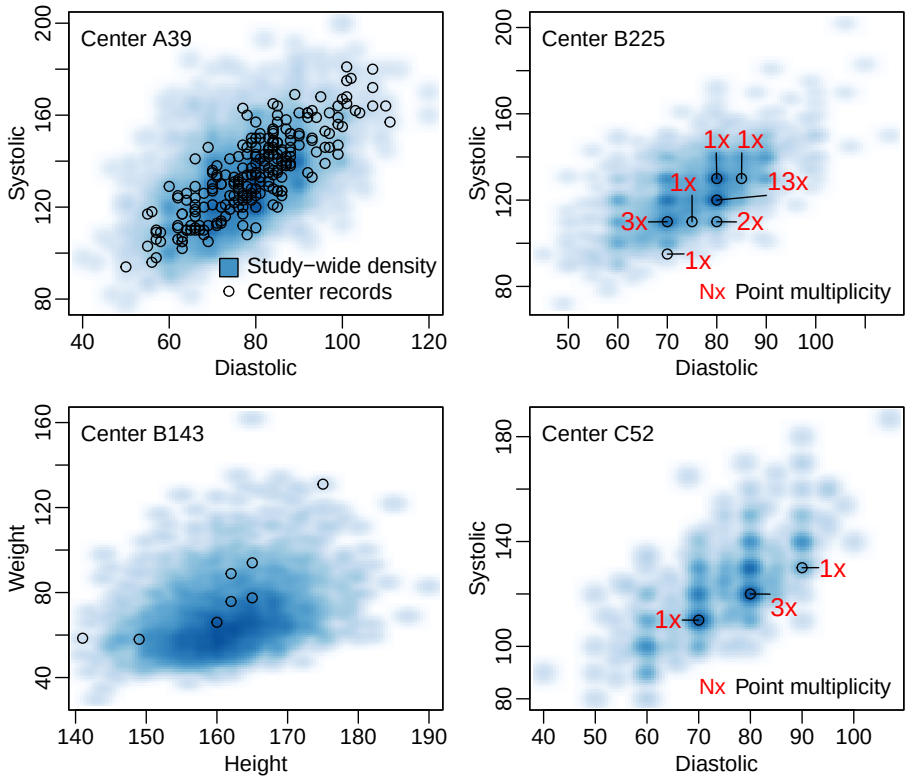

FIGURE 9 Four illustrative cases that were detected by one or both tests. Center data (black) are contrasted with overall cloud (blue density). If duplicated points are present, their count is displayed.Center A39 was flagged by both methods. Center B225 was flagged only by the fixed margin test as it had low marginal variances, making its corrlation more significant. Center B143 was flagged only by the Fisher scale test as it had very large marginal variances, explaining the high correlation. Center C52 illustrates the problem of discretization, in this case leading to a perfect correlation.

and center-to-center variability. There are two main costs associated with using the margin restricted method. First, two tests are performed on each pair (conditioning on each of the two features) and their results do not necessarily agree, which can make interpretation difficult. It also implies a small loss in power, when correcting for multiple testing. Second, the expected correlation varies from center to center. It is possible that the same correlation may be considered as significantly too high in a center with a very small standard deviation, and too low in a center with a very large standard deviation. In a way this is reality, and from a CSM point of view both situations may be considered as suspect, but it can make reporting and interpretation of the results challenging. Another difference between the two methods is that the Fisher test, being based on correlations, is independent of the scale of the variables in individual centers, while the fixed margin method uses the scale of one feature to estimate the expected center correlation. This makes the fixed margin method unsuitable for cases where correlations are expected to be similar across centers, but each center may use a different scale. This can happen for some blood measurements for instance, for which the scale may be dependent on the laboratory apparatus.

It is important to stress that the tests presented are meant to be part of a suite of tests for CSM (as described in the Introduction). Those other tests could analyse differences in means, variances, and other parameters of interest. Combining all those tests together dramatically increases the potential to find outliers. This can be readily illustrated with the example discussed in Section 5.1. In this example, when the fixed margin test was combined with 2 tests of the marginal means and 2 tests of the marginal variances, we were able to detect almost all of the 34 fabricated datasets (in 31 we had a sensitivity of at least 90% instead of in about half when the detection was based solely on the fixed margin test). These are the conclusions of an analysis reported in a separate manuscript[20]. Indeed, when data are fabricated in one or a few centers, this may also be visible on the univariate level in terms of discrepancies in the means or standard deviations of the variables X or Y across centers.

Since in CSM we pool information (in particular p values) from several tests, the properties of these individual tests are important and must be similar. Robust specificity in any individual test is crucial to prevent false alarms at the CSM level, while a lower sensitivity in a single test can be overcome by the fact that several other tests contribute to the signal when the data present some anomaly.

A recent scoping review of the broader field of *risk-based monitoring* of clinical trials (as opposed to extensive source data verification), and in which CSM plays a key role, is given in Cragg et al.[21].

It is important to point out that detection of outlyingness with our methods does certainly not imply that a center is fraudulent, and does not even necessarily point to a quality issue. While acknowledging the progress in AI systems, an automated method, even carefully designed, cannot completely replace the analyst and the thorough investigations required to make sound judgments regarding fraud. Therefore, significant results of tests are considered merely as signals that are worth looking into. The philosophy of the CSM approach is that a systematic issue with a center or patient will be reflected simultaneously in multiple statistical tests and hence should contribute to the summarizing score, while a spurious phenomenon, due to the play of chance, should be more easily filtered out when a large number of tests are combined. Nevertheless, interpretation should not be handled by the system in an automated way, but rather be supervised by an analyst.

We add that future extensions of the tests can be considered. A result similar to the approximation given in (3) may be obtained for the Spearman correlation coefficient r_S, namely an approximately normal model for $z(r_S)$ with variance $1.06/(n-3)$. This opens the way for the Fisher method to be applied to a larger class of models. More details on this empirical result are given in[22]. However, since the result is obtained under the assumption of a bivariate normal for the data, this extension encompasses models where both variables are a monotonic transformation of the margins in such model. This could be a possible way to address non-normal data. Unfortunately, to the best of our knowledge, the fixed margin approach does not seem to allow such extension.

ACKNOWLEDGMENTS

The authors thank Drs Akhtar-Danesh and Dehghan-Kooshkghazi for providing the datasets of fabricated data used in Section 5.1 of this paper.

Software availability statement

The code of the methods presented is available as the `CSMtests` R package on the public repository https://github.com/dvenet/CSMtests.

Data availability statement

The distributions of center sizes are available in the package. The data that support the findings are only partially available. For the fabricated data, on request to Dr Noori Akhtar-Danesh. The clinical data are not publicly available due to privacy, legal or ethical restrictions.

Financial disclosure

None reported.

Conflict of interest

LT is an employee and shareholder of CluePoints. TB and MB are employees and shareholders of IDDI. MB is a shareholder of CluePoints. DV receives fees from CluePoints.

References

1. Buyse M, George S, Evans S, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Stat Med* 1999; 18(24): 3435-52.

2. George SL, Buyse M. Data fraud in clinical trials. *Clinical investigation* 2015; 5(2): 161.

3. Akhtar-Danesh N, Dehghan-Kooshkghazi M. How does correlation structure differ between real and fabricated data-sets?. *BMC Med Res Method* 2003; 3(18).

4. Venet D, Doffagne E, Burzykowski T, et al. A statistical approach to central monitoring of data quality in clinical trials. *Clin Trials* 2012; 9(6): 705-13.

5. Trotta L, Kabeya Y, Buyse M, et al. Detection of atypical data in multicenter clinical trials using unsupervised statistical monitoring. *Clinical Trials* 2019; 16(5): 512–522.

6. Buyse M, Trotta L, Saad ED, Sakamoto J. Central statistical monitoring of investigator-led clinical trials in oncology. *International Journal of Clinical Oncology* 2020; 25(7): 1207–1214.

7. Desmet L, Venet D, Doffagne E, et al. Linear mixed-effects models for central statistical monitoring of multisite clinical trials. *Stat Med* 2014; 33: 5265-79.

8. Desmet L, Venet D, Doffagne E, et al. Use of the beta-binomial model for central statistical monitoring of multicenter clinical trials.. *Stat Biopharm Res* 2017; 9(1): 1–11.

9. Xu J, Huang L, Yao Z, Xu Z, Zalkikar J, Tiwari R. Statistical methods for clinical study site selection. *Therapeutic Innovation & Regulatory Science* 2020; 54: 211–219.

10. Viron dS, Trotta L, Schumacher H, et al. Detection of fraud in a clinical trial using unsupervised statistical monitoring. *Therapeutic Innovation & Regulatory Science* 2022; 56: 130–136.

11. Bor v. dRM, Vaessen PW, Oosterman BJ, Zuithoff NP, Grobbee DE, Roes KC. A computationally simple central monitoring procedure, effectively applied to empirical trial data with known fraud. *Journal of Clinical Epidemiology* 2017; 87: 59–69.

12. Larntz K, Perlman M. A Simple Test for the Equality of Correlation Matrices. In: Gupta S, Berger J., eds. *Statistical Decision Theory and Related Topics IV*. 2. Springer-Verlag; 1988; New York: 289–298.

13. Koziol J, Alexander J, Bauer L, et al. A graphical technique for displaying correlation matrices. *Am Statistician* 1997; 51: 301–304.

14. Taylor RN, McEntegart DJ, Stillman EC. Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Information Journal* 2002; 36: 115–125. doi: 10.1177/009286150203600115

15. Fisher R. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika* 1915; 10(4): 507–521.

16. Gayen A. The Frequency Distribution of the Product-Moment Correlation Coefficient in Random Samples of Any Size Drawn from Non-Normal Universes. *Biometrika* 1951; 38(1): 219-47.

17. Diedenhofen B, Musch J. cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE* 2015; 10(4): e0121945. doi: 10.1371/journal.pone.0121945

18. Kendall M. *The Advanced Theory of Statistics. Volume 1 (second edition)*. London, UK: Charles Griffin & Co . 1945.

19. Hogben D. The Distribution of the Sample Correlation Coefficient With One Variable Fixed. *J Res Nat Bur Stand* 1968; 72B(1): 33-35.

20. Desmet L, Venet D, Akhtar-Danesh N, Trotta L, Burzykowski T, Buyse M. Detection of data fabrication in multicenter experiments. *Personal Communication* 2021.

21. Cragg WJ, Hurley C, Yorke-Edwards V, Stenning SP. Dynamic methods for ongoing assessment of site-level risk in risk-based monitoring of clinical trials: A scoping review. *Clinical Trials* 2021; 18(2): 245–259.

22. Fieller E, Hartley H, Pearson E. Tests for Rank Correlation Coefficients. *Biometrika* 1957; 44(3/4): 470–481.

How to cite this article: L. Desmet, D. Venet, L. Trotta, T. Burzykowski, and M. Buyse (2024), Detection of outlying correlation coefficients in multicenter clinical trials, *Pharmaceutical Statistics*, *2024;00:1–6*.