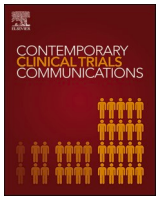
Contemporary Clinical Trials Communications 34 (2023) 101168

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc



Performance of four centralized statistical monitoring methods for early detection of an atypical center in a multicenter study

Serge Niangoran [a,b,c,*], Valérie Journot [a,b], Olivier Marcy [a,b], Xavier Anglaret [a,b,1], Amadou Alioum [a,1]

[a] University of Bordeaux, National Institute for Health and Medical Research (INSERM) UMR 1219, Bordeaux Population Health Research Center, Bordeaux, France
[b] Research Institute for Sustainable Development (IRD) EMR 271, Bordeaux, France
[c] Programme PACCI, Abidjan, Côte d'Ivoire

ARTICLE INFO

Keywords:
Data quality
Centralized Statistical monitoring
Multicenter clinical trial
Sensitivity
Specificity

ABSTRACT

Background: Ensuring the quality of data is essential for the credibility of a multicenter clinical trial. Centralized Statistical Monitoring (CSM) of data allows the detection of a center in which the distribution of a specific variable is atypical compared to other centers. The ideal CSM method should allow early detection of problem and therefore involve the fewest possible participants.
Methods: We simulated clinical trials and compared the performance of four CSM methods (Student, Hatayama, Desmet, Distance) to detect whether the distribution of a quantitative variable was atypical in one center in relation to the others, with different numbers of participants and different mean deviation amplitudes.
Results: The Student and Hatayama methods had good sensitivity but poor specificity, which disqualifies them for practical use in CSM. The Desmet and Distance methods had very high specificity for detecting all the mean deviations tested (including small values) but low sensitivity with mean deviations less than 50%.
Conclusion: Although the Student and Hatayama methods are more sensitive, their low specificity would lead to too many alerts being triggered, which would result in additional unnecessary control work to ensure data quality. The Desmet and Distance methods have low sensitivity when the deviation from the mean is low, suggesting that the CSM should be used alongside other conventional monitoring procedures rather than replacing them. However, they have excellent specificity, which suggests they can be applied routinely, since using them takes up no time at central level and does not cause any unnecessary workload in investigating centers.

1. Background

Clinical trials involving large numbers of participants can now be carried out more quickly at many clinical centers around the world due to the availability of increasingly efficient tools [1].

The quality of the data collected is essential for the credibility of the results of clinical trials [2]. To ensure quality, checks are carried out during collection of the data [3]. However, it is difficult to monitor all the data collected on site, and even more so to detect problems in time to resolve them [4]. Centralized statistical monitoring (CSM) of the database has been proposed to quickly identify one study center in which the distribution of a variable is atypical in relation to other centers, prompting action to confirm the problem and correct as necessary [5,6]. The ideal CSM method would enable the detection of problems as soon as possible after a new center is opened, and therefore has the fewest possible participants.

Several CSM methods have been proposed in the literature [7–9]. The simplest one is the Student's *t*-test comparing the variable mean in one site to the mean in all other sites. Other methods use more complex modeling, including those developed by Desmet et al. [10] using a linear mixed model approach, Hatayama and Yasui [11] using a Bayesian approach based on finite mixture models, and Pogue et al. [5] using the natural logarithm of the distance between the variable mean in one atypical center in relation to other centers [5,12,13].

In this paper, we first outline the basics of the Hatayama and Yasui [11] and Desmet et al. [10] methods, and we propose a new statistical

* Corresponding author. Inserm U1219, Université de Bordeaux, 146 rue Léo Saignat, 33076, Bordeaux, Cedex, France.
E-mail address: bessekon.niangoran@u-bordeaux.fr (S. Niangoran).
[1] These authors contributed equally.

https://doi.org/10.1016/j.conctc.2023.101168
Received 14 February 2023; Received in revised form 2 June 2023; Accepted 18 June 2023
Available online 29 June 2023

Table 1
Simulation parameters.

		Base case	Range
Sample size in the atypical center	N_a	50	10 to 300
Sample size in the overall study	N	N_a* 10	$N_a * 4$ *to* $N_a * 20$
Continuous variable Y			
Y mean in non-atypical centers	$\bar{y}_{na.}$	10	10 to 10,000
Mean shift in the atypical center	$(\bar{y}_{a.} - \bar{y}_{na.})/\bar{y}_{na.}$	10%	10%–100%
Intracenter variance (all centers)	σ_s^2	1	–
Residual variance (all centers)	σ_ε^2	4	–

test, called the 'Distance method', for detecting atypical distribution, inspired by the work of Pogue et al. [5]. Then we describe how we simulated a study specifically built to compare the performance of the Student, Hatayama, Desmet and Distance methods in determining whether the distribution of a quantitative variable in a given site is atypical in relation to other sites. Finally, we present and discuss the results of the simulation study.

2. Methods

2.1. Basics of the CSM methods

In this section, we consider a Gaussian-assumed quantitative Y variable, collected in a multicenter study with M sites. The observed jth value ($j = 1, \ldots, N_i$) of the variable Y in site i ($i = 1, \ldots, M$) is denoted y_{ij}, and the mean of y_{ij} in site i is denoted $\bar{y}_{i.} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$.

2.1.1. Hatayama Method [11]

This method uses a Bayesian finite mixture model approach. A model based on a finite mixture of probability distributions assumes, by definition, that the observed data set is from a source containing several homogeneous subpopulations, called components. The term 'mixture' therefore refers to the underlying assumption that the observed data is not generated from a single probability distribution but is sampled from K probability distributions (K > 1). The probability density of the mixture distribution is consequently written as follows:

$$f(y|\Theta) = \sum_{k=1}^{K} \pi_k f_k(y|\theta_k) \quad (1)$$

where:

- f_k is the probability density function of the kth component.
- $y = (y_1, y_2, \ldots, y_q)$ is the q-vector of the values of the quantitative variable of interest.
- π_k represents the k component mixture proportion, with $0 < \pi_k \leqslant 1$ and $\sum_{k=1}^{K} \pi_k = 1$.
- θ_k is the unknown parameter specific to the kth component distribution of the mixture.

In this approach, the data set is taken from a mixture of distributions, the components of which are distributions for atypical and non-atypical sites. With Gaussian mixture models, the values y_{ij} for participants in the non-atypical centers follow a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, and the values y_{ij} for participants in the atypical centers are also observed as being Gaussian $\mathcal{N}(\mu + \Delta, \sigma^2)$. Using moments of order 1 and 2 for estimating the expectation and variance of the mixture model, observations y_{ij} resulting from the mixture of the two types of site, therefore, have the following distribution:

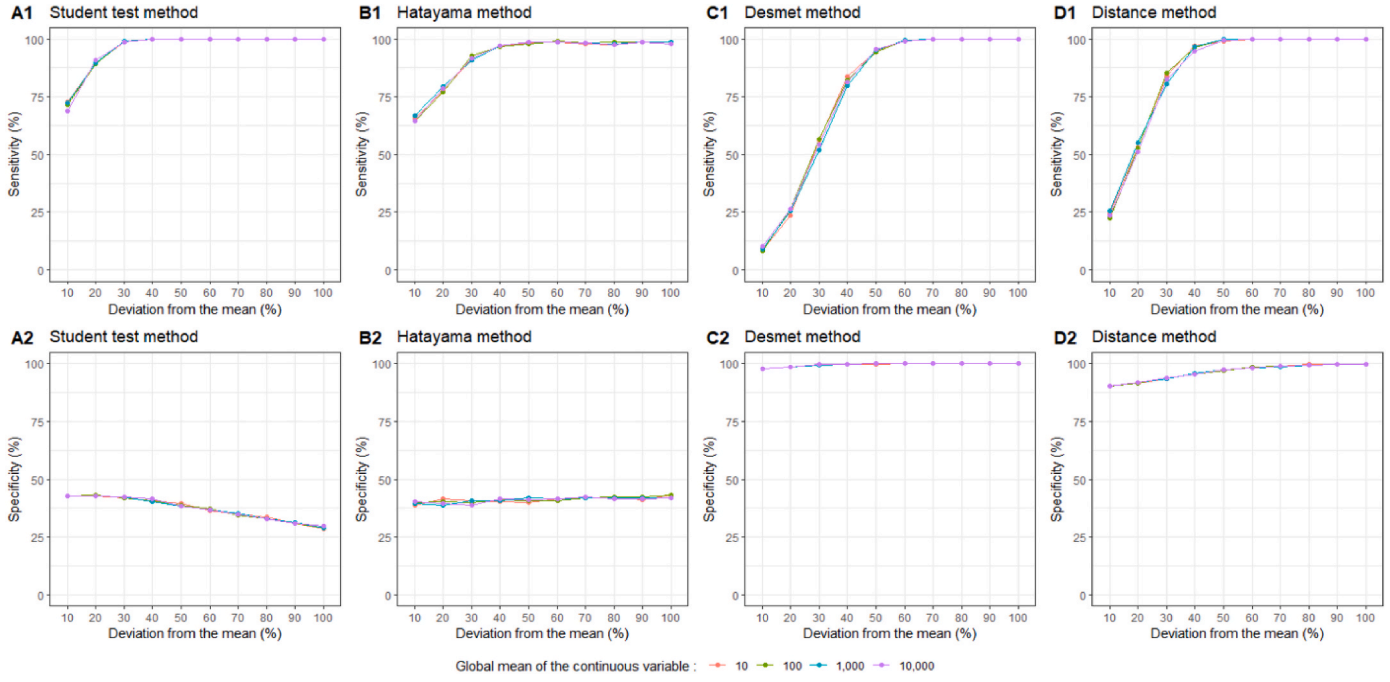


Fig. 1. Sensitivity and specificity of each centralized statistical monitoring method to detect the atypical center under base case scenario
Footnotes to Fig. 1
This figure explores the sensitivity and specificity of each of the four methods to detect that a center is atypical in the distribution of a continuous variable Y, in a simulated trial including 10 centers with the same number of participants (50) in each center.
The results are shown with:
• $(\bar{y}_{a.} - \bar{y}_{na.})/\bar{y}_{na.}$ varying from 10 to 100% (horizontal axis); With $\bar{y}_{na.}$ = mean of the Y values in the non-atypical centers and $\bar{y}_{a.}$ = mean of the Y values in the atypical center.
• $\bar{y}_{na.}$ absolute values of 10, 100, 1000 and 10,000 (coloured curves), to ensure that the model is robust and shows similar results irrespective of the absolute value of the continuous variable studied.
In these simulations, the ratio N_a/N remains constant (1/10).

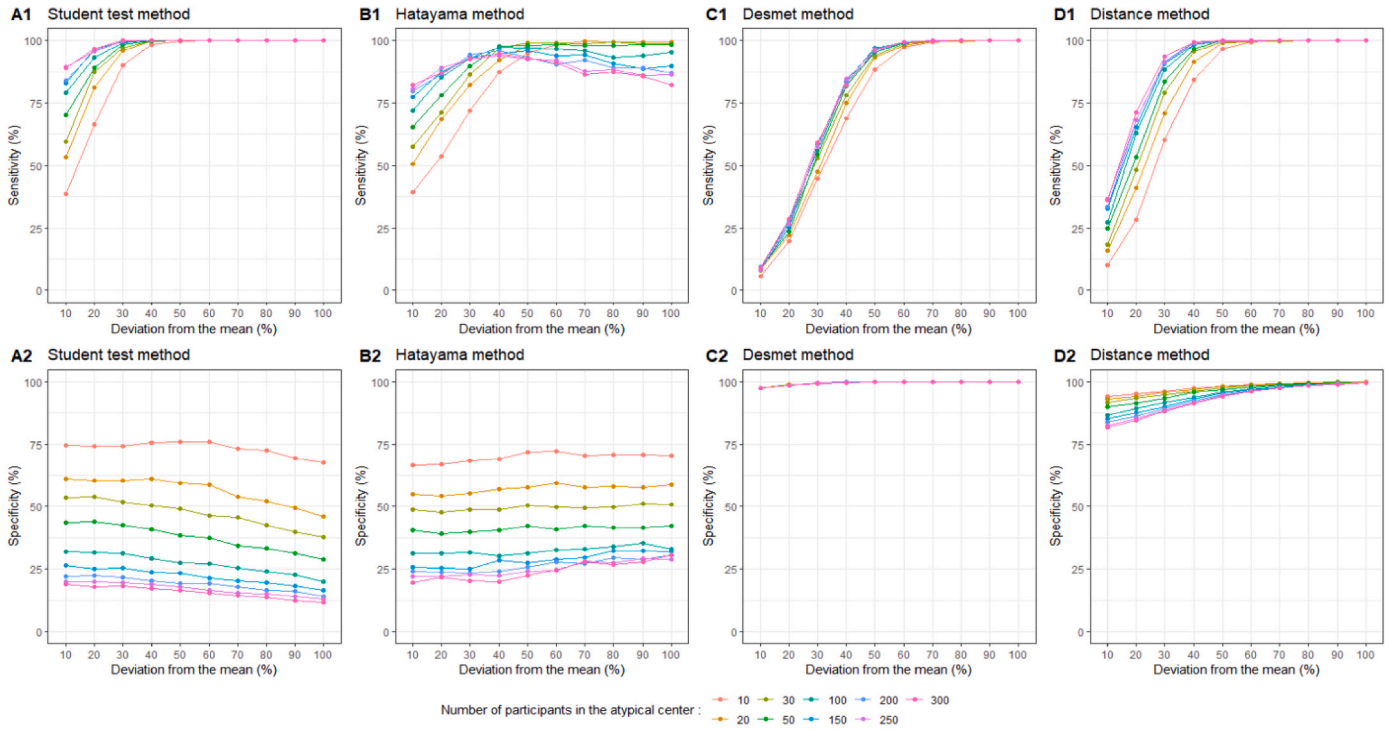


Fig. 2. Sensitivity and specificity of each centralized statistical monitoring method to detect the atypical center when different number of participants are included in the atypical center

Footnotes to Fig. 2

This figure explores the sensitivity and specificity of each of the four methods to detect that a center is atypical in the distribution of a continuous variable Y, in a simulated trial including 10 centers with the same number of participants in each center.

The results are shown with:

• $(\overline{y}_{a.}\text{-}\overline{y}_{na.})/\overline{y}_{na.}$ varying from 10 to 100% (horizontal axis); With $\overline{y}_{na.}$ = mean of the Y values in the non-atypical centers and $\overline{y}_{a.}$ = mean of the Y values in the atypical center.

• the number of participants per center varying from 10 to 300 (coloured curves).

In these simulations, the ratio N_a/N remains constant (1/10).

$$y_{ij} \sim \mathcal{N}\left(\mu + r\Delta, \sigma^2 + r(1-r)\,\Delta^2\right) \text{ with i} = 1, \ldots \text{M and j} = 1, \ldots, N_i \qquad (2)$$

where r ($0 < r < 1$) is the ratio of participants from atypical sites over all participants.

Accordingly, the mean $\overline{y}_{i.} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$ follows a Gaussian distribution:

$$\overline{y}_{i.} \sim \mathcal{N}\left(\mu + r\Delta, \frac{\sigma^2 + r(1-r)\,\Delta^2}{N_i}\right)$$

Assuming a finite mixture of Gaussian models of data from atypical and non-atypical centers, the Bayesian statistical model is summarized by the formulation of likelihood

$$\prod_{i=1}^{M}\prod_{j=1}^{N_i}\left(\sum_{k=1}^{K}\pi_k f\left(y_{ij}\middle|\boldsymbol{\theta_k}\right)\right)$$

where $f(y_{ij}\middle|\boldsymbol{\theta_k})$ is the density of a Gaussian distribution $\mathcal{N}(\mu_k, \sigma_k)$. The parameters π_k, μ_k, σ_k are estimated from all the data using a Bayesian approach by specifying prior distributions for these parameters [11].

The majority distribution indicated by $k_b = \text{argmax}\, E(\pi_k|Y)$ (also called the "Body distribution") is then used to calculate the posterior predictive distributions of the means of each center and the quantiles needed for center evaluation, using Equation (3).

$$p_{k_b}(\overline{y}_{i.}|Y) = \int f\left(\overline{y}_{i.}\middle|\theta_{k_b}\right)p\left(\theta_{k_b}\middle|\,Y\right)d\theta_{k_b} \qquad (3)$$

where $p(\theta_{k_b}\middle|\,Y)$ is the a posteriori density function of parameter θ_{k_b} of the body distribution, knowing the set of observations Y.

By choosing a risk α, a decision rule can be used to define the critical region for detecting atypical sites. A given site i will therefore be considered atypical if $\overline{y}_{i.} \notin [\gamma_i^{\alpha}\text{-}\gamma_i^{1-\alpha}]$, where $\overline{y}_{i.}$ is the observed mean and γ_i^{α} is the $100\ \alpha - th$ percentile of the $p_{k_b}(\overline{y}_{i.}|Y)$.

2.1.2. *Desmet Method [10]*

This method is based on a hybrid model which combines data from two normal distributions. It assumes a continuous variable with a Gaussian distribution, with two subsets of observations. The first subset of size n_0, mean μ_0 and standard deviation σ is called the null model, and corresponds to a normal model followed by the majority of the observations. The second subset of size n_1, mean μ_1 and the same standard deviation σ is called the alternative model, and contains the data whose mean is shifted with respect to the null model; μ_1 is assumed to be equal to $\mu_0 + \delta$, where δ can be positive or negative. The data resulting from the fusion of these two distributions (the hybrid model) show normal distribution $\mathcal{N}(\mu_{hybrid}, \sigma^2_{hybrid})$. The hybrid model is a good approximation of the null model when n_1 is sufficiently small compared to n_0.

$$\text{Desmet et al. proposed a linear mixed} - \text{effects model} : y_{ij} = \mu + \gamma_i + \varepsilon_{ij} \qquad (4)$$

$$\text{with } \gamma_i \ i.i.d. \sim \mathcal{N}\left(0, \sigma_s^2\right) \text{ and } \varepsilon_{ij}\ i.i.d. \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$$

where γ_i is the random effect for the site, i.e. the variability linked to the site, ε_{ij} the random residual error, σ_s^2 the within-center variance and σ_ε^2 the residual variance.

Under the assumption of the model, the mean $\overline{Y}_{i.}$ in site i follows a

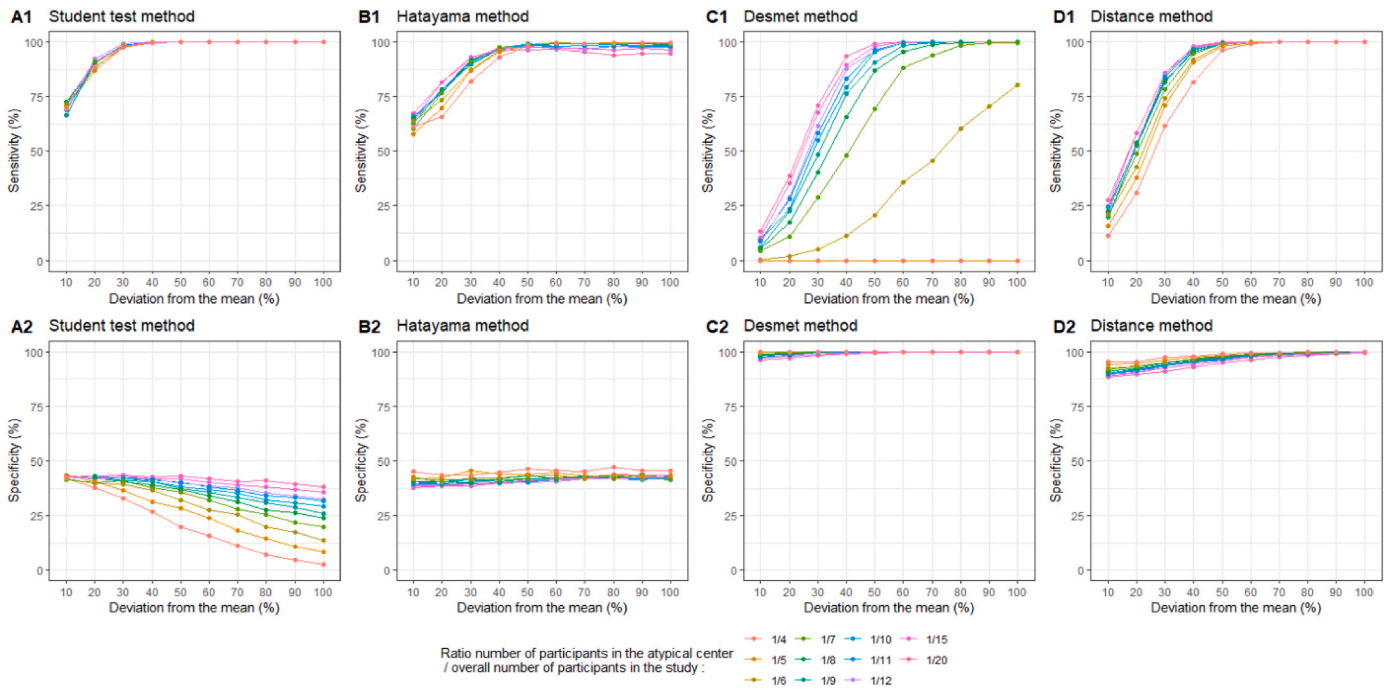


Fig. 3. Sensitivity and specificity of each centralized statistical monitoring method to detect the atypical center when varying the ratio N_a/N of the number of participants in the atypical center on the overall number of participants in the study

Footnotes to Fig. 3

This figure explores the sensitivity and specificity of each of the four methods to conclude that a center is atypical in the distribution of a continuous variable Y in a simulated trial including a varying number of centers with the same number of participants in each center.

The results are shown with:

• $(\bar{y}_{a.}\text{-}\bar{y}_{na.})/\bar{y}_{na.}$ varying from 10 to 100% (horizontal axis); With $\bar{y}_{na.}$ = mean of the Y values in the non-atypical centers and $\bar{y}_{a.}$ = mean of the Y values in the atypical center.

• The number of trial centers varies from 4 to 20, and therefore the ratio N_a/N varies from ¼ to 1/20 (coloured curves).

$\mathcal{N}\left(\mu, \sigma_s^2 + \frac{\sigma_\varepsilon^2}{Ni}\right)$ and can be used to detect atypical sites. The parameters μ, $\sigma_s^2, \sigma_\varepsilon^2$ are unknown but can be estimated from the linear mixed-effects model using all the data as the hybrid model to obtain $\widehat{\mu}_{hybrid}, \widehat{\sigma}_s^2, \widehat{\sigma}_\varepsilon^2$.

For each site i, we can assign a p-value using the statistic test

$$U_i = \frac{\overline{Y}_i - \widehat{\mu}_{hybrid}}{\sqrt{\widehat{\sigma}_s^2 + \frac{\widehat{\sigma}_\varepsilon^2}{N_i}}} \sim \mathcal{N}(0,1)$$

by applying the following rule

- If the calculated value u_i of U_i is negative , that is if $\overline{y}_{i.} \leq \widehat{\mu}_{hybrid}$, then the p-value is $p(\overline{y}_{i.}) = 2P(U_i < u_i)$.
- If the calculated value u_i of U_i is positive, that is if $\overline{y}_{i.} > \widehat{\mu}_{hybrid}$, then the p-value is $p(\overline{y}_{i.}) = 2P(U_i > u_i)$.

For a fixed threshold α, a center *i* is considered atypical only if the p-value $p(\overline{y}_{i.}) < \alpha$. This means that atypical centers are those with means located at the tail of the distribution of the hybrid model, delimited by the α/2 and 1-α/2 quantiles for a fixed α value.

2.1.3. Distance method

Pogue et al. [5] defined the distance d_i that measures how far away the data of center *i* are from the overall mean across all centers ($\overline{y}$), standardized by the overall standard deviation(s) by:

$$d_i = \sum_{j=1}^{N_i} \left(\frac{y_{ij} - \overline{y}}{s}\right)^2$$

where $\overline{y} = \frac{1}{\sum_{i=1}^{M} N_i} \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij}$ and $s^2 = \frac{1}{\left(\sum_{i=1}^{M} N_i\right) - 1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left(y_{ij} - \overline{y}\right)^2$

They used the natural logarithm of this distance as a predictor in models for detecting fraud and other systematic data irregularities in clinical trials [5].

Note that this distance as defined by Pogue et al. does not follow a particular theoretical distribution. However, by rewriting d_i in the form

$$d_i = \frac{\sum_{j=1}^{N_i} \left(y_{ij} - \overline{y}\right)^2}{s^2}$$

it was observed that, dividing the numerator by the degree of freedom $(N_i - 1)$, the quantity:

$$D_i = \frac{\frac{1}{(N_i - 1)} \sum_{j=1}^{N_i} \left(y_{ij} - \overline{y}\right)^2}{\frac{1}{\left(\sum_{i=1}^{M} N_i\right) - 1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left(y_{ij} - \overline{y}\right)^2} = \frac{d_i}{(N_i - 1)}$$

can be expressed as the ratio of two variances, which follows a Fisher-Snedecor distribution with degrees of freedom $\mathrm{df1} = (N_i - 1)$ and $\mathrm{df2} = (\sum_{i=1}^{M} N_i) - 1$: $D_i \sim F(\mathrm{df1}, \mathrm{df2})$.

Based on F-test in classical one-way analysis of variance (ANOVA) for the comparison of means, we had the idea to propose in this paper to use the quantity D_i as a statistic test to detect atypical site according to the following rule:

- If $D_i > F_{(1-\alpha)}(\mathrm{df1}, \mathrm{df2})$, then center *i* is considered as atypical.

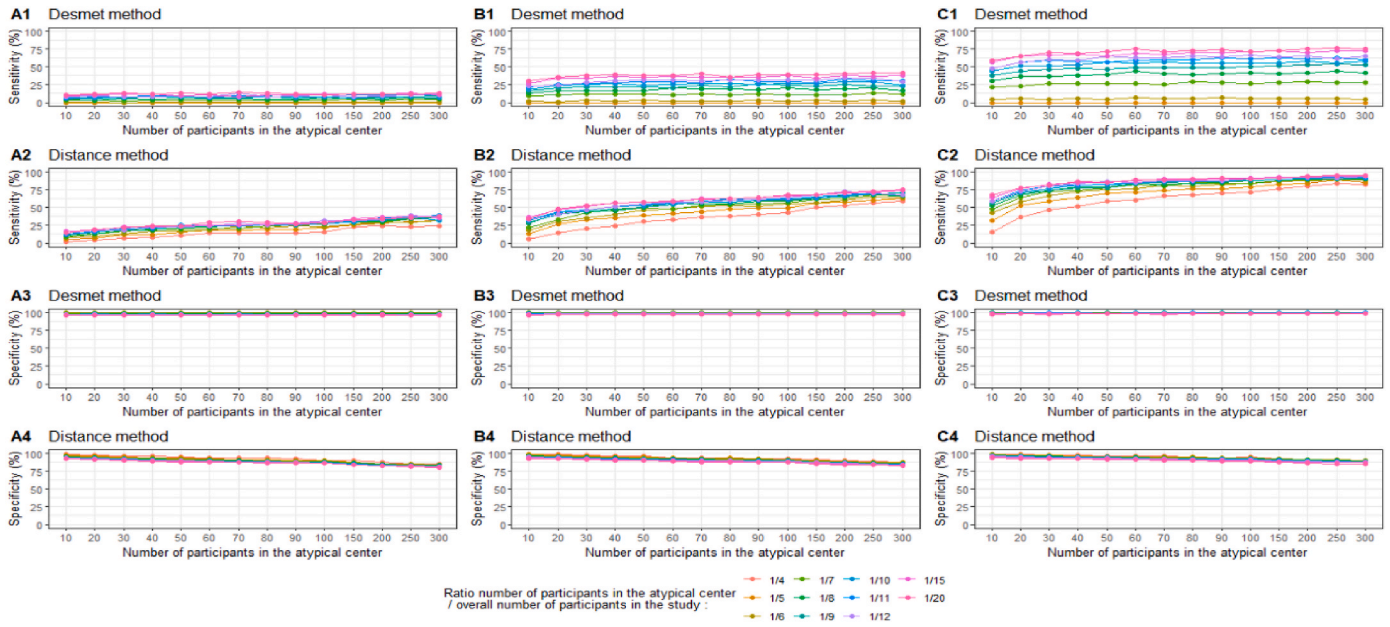


Fig. 4. Sensitivity and specificity of each centralized statistical monitoring method to detect low values of the deviation of the mean (10%, 20%, 30%), when different number of participants are included in the atypical center and when the ratio N_a/N varies

Footnotes to Fig. 4

This figure explores the sensitivity and specificity of the Desmet and Distance methods to conclude that a center may be atypical in the distribution of a continuous variable Y for low deviations of the mean and for different N_a and N_a/N ratios.

The results are shown with the number of participants in the atypical center N_a varying from 10 to 300 (horizontal axis) and the overall number of centers of the same size varying from 4 to 20 (coloured curves).

For each method, the sub-figure on the left (A) shows the result for $(\bar{y}_{a.}\text{-}\bar{y}_{na.})/\bar{y}_{na.} = 10\%$; the sub-figure on the middle (B) shows the result for $(\bar{y}_{a.}\text{-}\bar{y}_{na.})/\bar{y}_{na.} = 20\%$; and the sub-figure on the right (C) shows the result for $(\bar{y}_{a.}\text{-}\bar{y}_{na.})/\bar{y}_{na.} = 30\%$; With $\bar{y}_{na.}$ = mean of the Y values in the non-atypical study centers, and $\bar{y}_{a.}$ = mean of Y values in the atypical center *(a)*.

In these simulations, the ratio N_a/N varies from 1/4 to 1/20.

- If $D_i \leq F_{(1-\alpha)}(\mathrm{df1}, \mathrm{df2})$, then center i is considered as non-atypical.

Where $\boldsymbol{F_{(1-\alpha)}(df1, df2)}$ is the $(\mathbf{1}-\boldsymbol{\alpha}) \times \mathbf{100}\%$ centile of the Fisher distribution F(df1,df2) for a fixed risk α.

We can thus compare the performance in terms of sensitivity and specificity of this new distance method to three other existing CSM methods.

2.2. Simulation study

2.2.1. Overview

We simulated multicenter clinical trials to assess the sensitivity and specificity of four CSM methods to detect as early as possible whether the distribution of a continuous variable Y was "atypical" in one trial center in relation to the other centers. The four methods were the Student, Hatayama, Desmet and Distance methods.

We assumed that: (i) All trial centers had the same number of participants; (ii) Only one center had an "atypical" distribution of Y, regardless of the number of centers; (iii) The distribution of Y was Gaussian in both the "atypical" and "non-atypical" centers.

Each simulation was replicated 1000 times. The sensitivity and specificity of each method to detect the atypical center were computed by counting in each simulation the number of true positives (#TP), true negatives (#TN), false positives (#FP), and false negatives (#FN). Sensitivity and specificity were calculated as follows:

$$\text{sensitivity} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}}$$

$$\text{specificity} = \frac{\#\text{TN}}{\#\text{FP} + \#\text{TN}}$$

2.2.2. Data Generation and transformation

y_{ij} values were generated using a two-level hierarchical model: the sites were generated at the first level and the participants in each site at the second level (equation (4)). Data was generated in one atypical center and a number of non-atypical centers. The mean value of Y was similar in all non-atypical centers ($\bar{y}_{na.}$). In the atypical center, the mean value of Y was $\bar{y}_{a.} = \bar{y}_{na.} + \delta$. The deviation of the mean between the atypical center and the other centers was expressed as a percentage $[(\bar{y}_{a.}\text{-}\bar{y}_{na.})/\bar{y}_{na.}]$.

To eliminate any effect that absolute values of Y could have on the performance of the CSM methods, we transformed the y_{ij} values with expectation $\mathrm{E}(y_{ij}) = \mu$ and variance $\mathrm{V}(y_{ij}) = \sigma_s^2 + \sigma_\varepsilon^2$ into centered-reduced values z_{ij}, using the following formula:

$$z_{ij} = \frac{y_{ij} - \mathrm{E}\left(y_{ij}\right)}{\sqrt{\mathrm{V}(\bar{y}_{i.})}} \tag{5}$$

such that $\bar{z}_{i.} = \frac{1}{N_i}\sum_{j=1}^{N_i} z_{ij} \sim \mathcal{N}(0,1)$

2.2.3. Scenarios

We first simulated a set of trials with 10 centers, in which the number of participants per center was 50 and the deviation of the mean between the atypical center and the other centers $[(\bar{y}_{a.}\text{-}\bar{y}_{na.})/\bar{y}_{na.}]$ varied between 10% and 100%. To estimate the robustness of the model and its ability to give identical results regardless of the absolute value of the mean of the untransformed variable Y, $\bar{y}_{na.}$ was also varied from 10 to 10,000.

After this first set of analysis, the number of participants in the atypical center N_a was varied from 10 to 300, keeping a number of centers at 10 (and therefore a ratio N_a/N at 1/10). Then, the number of

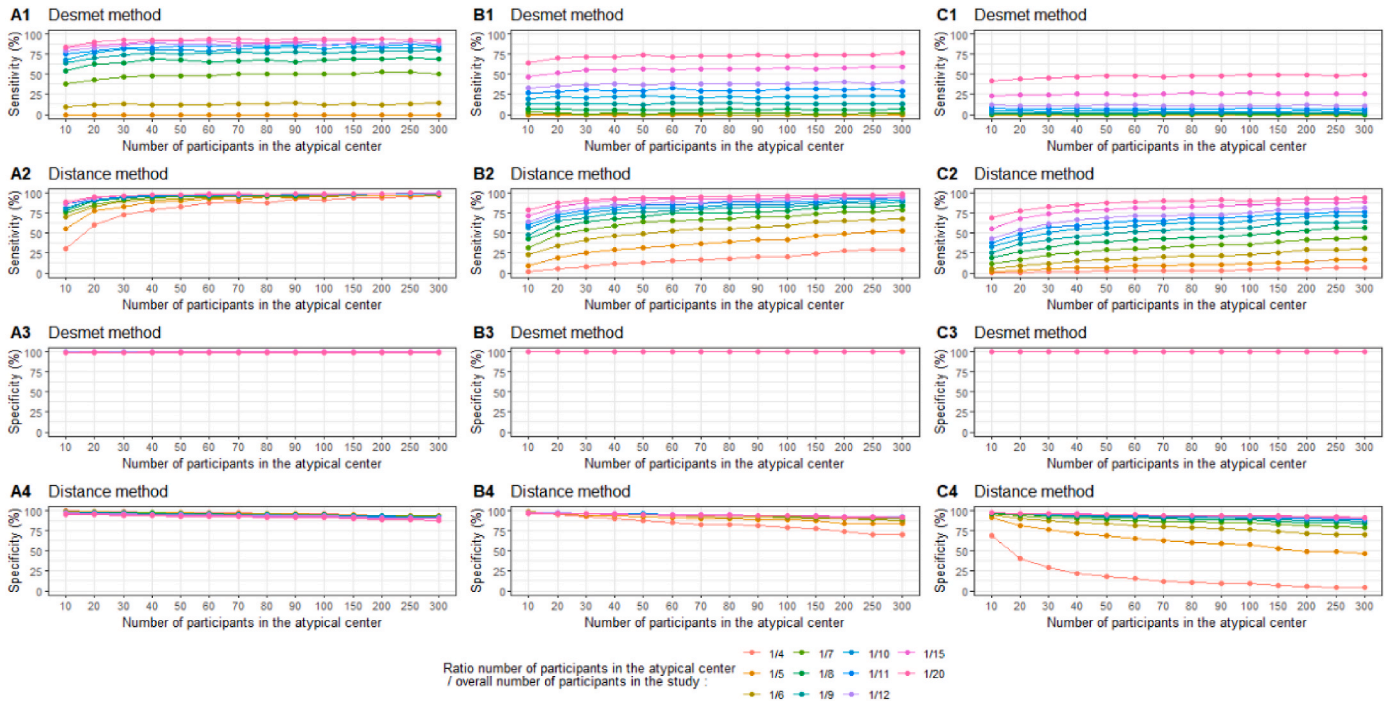


Fig. 5. Sensitivity and specificity of each centralized statistical monitoring method to detect an atypical center when other centers are atypical

Footnotes to Fig. 5

This figure explores the sensitivity and specificity of the Desmet and Distance methods to conclude that a center may be atypical in the distribution of a continuous variable Y in the scenario where there are other atypical centers.

The results are shown with the number of participants in the atypical center N_a varying from 10 to 300 (horizontal axis) and the number of centers of the same size varying from 4 to 20 (coloured curves).

All analyses are performed with $(\bar{y}_{a.} - \bar{y}_{na.})/\bar{y}_{na.} = 40\%$. With $\bar{y}_{na.}$ = mean of the Y values in the non-atypical study centers, and $\bar{y}_{a.}$ = mean of the Y values in the atypical center *(a)*.

For each method, the sub-figure on the left (A) shows the result when the only atypical center is that being analysed; the sub-figure on the middle (B) shows the results when there is another atypical center with the same number of participants; the sub-figure on the right (C) shows the results when there is another atypical center with twice the same number of participants.

In these simulations, the ratio N_a/N varies from 1/4 to 1/20.

centers was varied from 4 to 20 (and therefore the ratio N_a/N from 1/4 to 1/20), keeping the number of participants N_a in the atypical center at 50. Both N_a and N_a/N were then varied simultaneously (see Table 1).

Finally, the consequences of atypicality contamination were explored, assuming successively that: among the supposed "non atypical" there was another "atypical" center with the same number of participants as the atypical center on which the analysis was focused; among the supposed "non atypical" centers there were two other "atypical" centers, each with the same number of participants as the atypical center on which the analysis was focused.

All programs and functions were carried out using the R software (version 4.2.1) and the simulations were carried out on the computing clusters of the Aquitain Intensive Computing Mesocenter (MCIA).

We choose three components (K = 3) for Bayesian modelling of finite mixtures, as proposed by Hatayama.

For the implementation of the Hatayama method, we used Markov Chain Monte Carlo (MCMC) methods for the estimation of the parameters of the finite mixture model. To do this, we used the ***jags*** function of the ***jagsUI*** package under the R software.

3. Results

Fig. 1 shows the results of the base case analysis. With 50 participants in the atypical center and 500 participants across all centers (i.e. Na/N ratio = 1/10), sensitivity in detecting atypicality reached or exceeded 95% for each of the four methods where deviation of the mean in the atypical center reached or exceeded 50%. For a deviation of 40%, the Student method showed the best sensitivity (100%), followed by the Hatayama and Distance methods (both 97%), far ahead of the Desmet method (81%). For a deviation of 30%, the sensitivity of the Student method (98%), and the Hatayama method (87%) remained high, while that of the Distance method (83%) and Desmet method (54%) fell steadily. For deviations of 10% and 20%, the Distance method (24% and 52%) and Desmet method (7% and 26%) showed low sensitivity, while the Student method (69% and 90%), and the Hatayama method (61% and 73%) still performed well. However, the specificity of the Student and Hatayama methods never exceeded 45%, while that of the Desmet and Distance methods always exceeded 90% and tended towards 100% with increasing levels of mean deviation. The Desmet method had higher specificity than the Distance method for mean deviations ranging from 10% to 50%. The specificity of both the Desmet and Distance methods was close to 100% when the mean deviation exceeded 50%. All these results were similar whatever the absolute value of the mean of the variable studied.

Fig. 2 shows the influence of decreasing (from 50 to 10) or increasing (from 50 to 300) the number of participants in the atypical center, while keeping the ratio N_a/N at 1/10. Increasing the number of participants in the atypical center while keeping N_a/N constant resulted in increased sensitivity and decreased specificity in all tests where deviation of the mean was equal to or below 50%. For higher deviations, the specificity of the Student and Hatayama methods also decreased with fewer participants in the atypical center.

Fig. 3 shows the influence of increasing the ratio N_a/N (from 1/10 to 1/4) or decreasing it (from 1/10 to 1/20), while keeping N_a at 50. Decreasing the N_a/N ratio while keeping the N_a constant resulted in increased sensitivity and decreased specificity in all tests where mean

deviation was equal to or below 50%. For higher deviations of the mean, the sensitivity of the Desmet method also decreased with an increased N_a/N ratio.

Figs. 4 and 5 only show the results for the Desmet and Distance methods.

Fig. 4 explores the influence of varying the number of participants in the atypical center (from 10 to 300) and the N_a/N ratio (from 1/4 to 1/20) simultaneously for low mean deviations (10%, 20% and 30%). With the Desmet method, sensitivity improved with a decreased N_a/N, but increasing the N_a had little influence irrespective of the N_a/N ratio. The sensitivity of the Desmet method never exceeded 75%. With the Distance method, the increase in sensitivity was intensified by both increasing the N_a and decreasing the N_a/N

(that is increasing the number of centers), reaching 90% with a combination of $N_a = 100$ and $N_a/N = 1/7$.

Fig. 5 explores the possibility that there are one or two atypical centers in addition to the atypical center to be detected. With one or two other atypical centers among the "other" centers, the sensitivity to detecting the atypical center on which the analysis was focused was decreased, unsurprisingly, especially for the Desmet method when the N_a/N was high, and for the Distance method when the N_a was low and/or the N_a/N was high.

4. Discussion and conclusions

In this paper, the performance of four CSM methods is compared for the early detection of an atypical center in multicenter trials. "Early detection" means applying the method to each new center when it still has relatively few participants, without anticipating either the higher number of participants that the center will reach later in the study, or the total number of participants already included in other centers at the time of the analysis.

Conceptually, the Desmet and Hatayama methods rely on the existence of several merged distributions and an approximation of the hybrid model by the majoritarian distribution [10,11], and the Distance and Student methods rely on a sole data distribution [5]. Our simulations show that in terms of performance, however, the Distance method is close to the Desmet method and the Hatayama method is close to the Student method.

The Desmet and Distance methods have low sensitivity overall for low mean-deviation values but very high specificity for detecting all deviations of the mean (including small values), and the Student and Hatayama methods have better sensitivity for low mean-deviation values but very low specificity for detecting all deviations of the mean. Increasing the number of participants in the atypical center, or increasing the ratio of the number of participants in the atypical center to the number of participants in the study, did not fundamentally alter the findings. Although the Student and Hatayama methods are more sensitive than the other two, their low specificity disqualifies them for practical use in centralized statistical monitoring. The profile would lead to too many alerts being triggered, which would result in additional unnecessary control work to ensure data quality.

The high specificity of the Desmet and Distance methods gives them more potential for practical use in CSM [14–16] and merits more detailed discussion of their use. Both methods are very specific but have weakness in sensitivity, being highly sensitive only for high deviations of the mean, which does not necessarily correspond to a relevant clinical situation [17,18]. For variables such as weight and hemoglobin, for example, it is difficult to conceive a measurement error that would result in a mean in one atypical center being 50% lower (or higher) than the mean in the other centers [19,20]. For detecting deviations of the mean between 10% and 50%, especially with a low number of participants in the atypical center and/or a high ratio of number of participants in the atypical center to number of participants in the study, the Distance method seems to show somewhat higher sensitivity than the Desmet method. This advantage is offset by a decrease in specificity when the number of participants in the atypical center is increased. For detecting mean deviations of less than 50%, the best compromise between sensitivity and specificity therefore seems to be the Distance method with a low number of participants in the center studied, and the Desmet method with a high number of participants. For practical use, however, two comments can be added: (i) as long as a CSM method gives a very specific result, poor sensitivity does not prevent its routine use. The purpose of CSM is to detect problems in order to introduce controls; it is not a substitute for traditional monitoring, it is an additional tool [21,22]. If a problem is detected by the CSM method and the result is specific, it saves time. If the problem is not detected by the CSM method, it can be detected by other monitoring actions; (ii) interpretation of the sensitivity and specificity parameters could suggest a solution combining two methods [6,23], starting with the method with the best sensitivity and confirming with the one with the best specificity.

The four methods differ in terms of both theoretical conceptualization and software implementation. Although the data used were generated in accordance with a two-level model, calculating the parameters (mean and standard deviation) in the Student, Hatayama and Distance methods does not take this hierarchical structure into account, unlike the Desmet method which does take it into account with its use of the random effects linear mixed model. In contrast to the Distance, Student and Hatayama methods, where the performance is theoretically not calculated, note also that the sensitivity and specificity of the Desmet method can theoretically be predicted by formulas. Using these formulas, Desmet et al. [10] showed that for arbitrarily large center sizes, the denominator of the test statistic decreases until it reaches the center variance, and the sensitivity of the method increases until it reaches its maximum value.

The main limitation of our work is that it is a simulation using artificially-generated data. To assess the value of applying these CSM methods in real-life conditions, they would need to be applied in several multicenter trials, and outcomes would need to be collected to judge their usefulness. These outcomes include acceptability, feasibility, time consumption, ability to detect a real problem, ability to improve the quality of the data, and time and money actually saved [24,25]. Another limitation is that the purpose of CSM is to detect an atypical distribution, not to conclude that this atypical distribution is the result of a problem. In a multicenter trial, distributions may differ for some variables because the population is different.

In conclusion, two CSM methods, the Desmet and Distance methods, showed theoretical strength which makes them eligible for real-life use in multicenter trials. They could be proposed for early use, for example every 10 or 20 new participants in each new center. Both methods have low sensitivity when the deviation from the mean is less than 50%, suggesting that the CSM should not be the only tool used for detecting atypicality, but should be used alongside other conventional monitoring procedures rather than replacing them. However, both methods have excellent specificity, which suggests they can be applied routinely, since using them takes up no time at central level and does not cause any unnecessary workload in investigating centers. The Distance method is a little more sensitive with low numbers of participants.

Availability of data and materials

The dataset generated and analysed during the current study is available from the corresponding author on reasonable request.

Funding

Serge Niangoran received a scholarship from the French Institut de Recherche pour le Développement (IRD) from January 2020 to December 2022.

Authors' contributions

Serge Niangoran performed the review, designed and interpreted the analyses of data, and prepared the manuscript for publication. Amadou Alioum and Xavier Anglaret contributed to the design and interpretation of the analyses, and substantively revised the manuscript. Valérie Journot and Olivier Marcy conceived the initial idea of the topic and helped critically revise the paper. All authors read and approved the final manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

None.

Abreviations

CSM	Centralized Statistical Monitoring
MCIA	Aquitain Intensive Computing Mesocenter

References

[1] S. Deering, M.M. Grade, J.K. Uppal, L. Foschini, J.L. Juusola, A.M. Amdur, C. J. Stepnowsky, Accelerating research with technology: rapid recruitment for a large-scale web-based sleep study, JMIR Res. Protoc. 8 (2019), e10974, https://doi.org/10.2196/10974.

[2] L. Houston, Y. Probst, P. Yu, A. Martin, Exploring data quality management within clinical trials, Appl. Clin. Inf. 9 (2018), https://doi.org/10.1055/s-0037-1621702, 072–081.

[3] B. Krishnankutty, S. Bellary, N.B.R. Kumar, L.S. Moodahadu, Data management in clinical research: an overview, Indian J. Pharmacol. 44 (2012) 168–172, https://doi.org/10.4103/0253-7613.93842.

[4] D.B. Fogel, Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review, Contemp. Clin. Trials Commun. 11 (2018) 156–164, https://doi.org/10.1016/j.conctc.2018.08.001.

[5] J.M. Pogue, P.J. Devereaux, K. Thorlund, S. Yusuf, Central statistical monitoring: detecting fraud in clinical trials, Clin. Trials Lond. Engl. 10 (2013) 225–235, https://doi.org/10.1177/1740774512469312.

[6] D. Venet, E. Doffagne, T. Burzykowski, F. Beckers, Y. Tellier, E. Genevois-Marlin, U. Becker, V. Bee, V. Wilson, C. Legrand, M. Buyse, A statistical approach to central monitoring of data quality in clinical trials, Clin. Trials 9 (2012) 705–713, https://doi.org/10.1177/1740774512447898.

[7] C. Baigent, F.E. Harrell, M. Buyse, J.R. Emberson, D.G. Altman, Ensuring trial validity by data quality assurance and diversification of monitoring methods, Clin. Trials Lond. Engl. 5 (2008) 49–55, https://doi.org/10.1177/1740774507087554.

[8] M. Buyse, S.L. George, S. Evans, N.L. Geller, J. Ranstam, B. Scherrer, E. Lesaffre, G. Murray, L. Edler, J. Hutton, T. Colton, P. Lachenbruch, B.L. Verma, The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials, Stat. Med. 18 (1999) 3435–3451, https://doi.org/10.1002/(sici)1097-0258(19991230)18:24<3435::aid-sim365>3.0.co;2-o.

[9] A.A. Kirkwood, T. Cox, A. Hackshaw, Application of methods for central statistical monitoring in clinical trials, Clin. Trials 10 (2013) 783–806, https://doi.org/10.1177/1740774513494504.

[10] L. Desmet, D. Venet, E. Doffagne, C. Timmermans, T. Burzykowski, C. Legrand, M. Buyse, Linear mixed-effects models for central statistical monitoring of multicenter clinical trials, Stat. Med. 33 (2014) 5265–5279, https://doi.org/10.1002/sim.6294.

[11] T. Hatayama, S. Yasui, Bayesian central statistical monitoring using finite mixture models in multicenter clinical trials, Contemp. Clin. Trials Commun. 19 (2020), 100566, https://doi.org/10.1016/j.conctc.2020.100566.

[12] S. Evans, Statistical aspects of the detection of fraud, in: S. Lock, F. Wells (Eds.), Fraud Misconduct Biomed. Res., 2, BMJ, London, 1996, pp. 226–239.

[13] R.N. Taylor, D.J. McEntegart, E.C. Stillman, Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data, Drug Inf. J. 36 (2002) 115–125, https://doi.org/10.1177/009286150203600115.

[14] S. Asgari, A. Mehrnia, M. Moussavi, Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine, Comput. Biol. Med. 60 (2015) 132–142, https://doi.org/10.1016/j.compbiomed.2015.03.005.

[15] S. Gao, B. Hu, X. Zheng, Y. Cao, D. Liu, M. Sun, B. Jiao, L. Wang, Gonyautoxin 1/4 aptamers with high-affinity and high-specificity: from efficient selection to aptasensor application, Biosens. Bioelectron. 79 (2016) 938–944, https://doi.org/10.1016/j.bios.2016.01.032.

[16] L. Trotta, Y. Kabeya, M. Buyse, E. Doffagne, D. Venet, L. Desmet, T. Burzykowski, A. Tsuburaya, K. Yoshida, Y. Miyashita, S. Morita, J. Sakamoto, P. Praveen, K. Oba, Detection of atypical data in multicenter clinical trials using unsupervised statistical monitoring, Clin. Trials 16 (2019) 512–522, https://doi.org/10.1177/1740774519862564.

[17] J. Martin Bland, DouglasG. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, Lancet 327 (1986) 307–310, https://doi.org/10.1016/S0140-6736(86)90837-8.

[18] J. Sawyer, Measurement and prediction, clinical and statistical, Psychol. Bull. 66 (1966) 178–200, https://doi.org/10.1037/h0023624.

[19] L.B. Guthrie, E. Oken, J.A. Sterne, M.W. Gillman, R. Patel, K. Vilchuck, N. Bogdanovich, M.S. Kramer, R.M. Martin, Ongoing monitoring of data clustering in multicenter studies, BMC Med. Res. Methodol. 12 (2012) 29, https://doi.org/10.1186/1471-2288-12-29.

[20] R. Olsen, A.R. Bihlet, F. Kalakou, J.R. Andersen, The impact of clinical trial monitoring approaches on data integrity and cost—a review of current literature, Eur. J. Clin. Pharmacol. 72 (2016) 399–412, https://doi.org/10.1007/s00228-015-2004-y.

[21] S.B. Love, V. Yorke-Edwards, S. Lensen, M.R. Sydes, Monitoring in practice – how are UK academic clinical trials monitored? A survey, Trials 21 (2020) 59, https://doi.org/10.1186/s13063-019-3976-1.

[22] V. Tantsyura, I.M. Dunn, K. Fendt, Y.J. Kim, J. Waters, J. Mitchel, Risk-based monitoring: a closer statistical look at source document verification, queries, study size effects, and data quality, Ther. Innov. Regul. Sci. 49 (2015) 903–910, https://doi.org/10.1177/2168479015586001.

[23] M. Buyse, L. Trotta, E.D. Saad, J. Sakamoto, Central statistical monitoring of investigator-led clinical trials in oncology, Int. J. Clin. Oncol. 25 (2020) 1207–1214, https://doi.org/10.1007/s10147-020-01726-6.

[24] J.R. Andersen, I. Byrjalsen, A. Bihlet, F. Kalakou, H.C. Hoeck, G. Hansen, H. B. Hansen, M.A. Karsdal, B.J. Riis, Impact of source data verification on data quality in clinical trials: an empirical post hoc analysis of three phase 3 randomized clinical trials, Br. J. Clin. Pharmacol. 79 (2015) 660–668, https://doi.org/10.1111/bcp.12531.

[25] L. Houston, A. Martin, P. Yu, Y. Probst, Time-consuming and expensive data quality monitoring procedures persist in clinical trials: a national survey, Contemp, Clin. Trials 103 (2021), 106290, https://doi.org/10.1016/j.cct.2021.106290.