Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Contemporary Clinical Trials 104 (2021) 106368

Contents lists available at ScienceDirect

Contemporary Clinical Trials

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

Risk-based centralized data monitoring of clinical trials at the time of COVID-19 pandemic

Most Alina Afroz [a], Grant Schwarber [b], Mohammad Alfrad Nobel Bhuiyan [b,*]

[a] Department of Electrical and Electronic Engineering, Begum Rokeya University, Rangpur, Bangladesh
[b] Department of Biostatistics, Medpace, Inc., 5375 Medpace Way, Cincinnati, OH 45227, USA

ARTICLE INFO

Keywords:
Centralize data monitoring
Risk based data monitoring

ABSTRACT

Objectives: COVID-19 pandemic caused several alarming challenges for clinical trials. On-site source data verification (SDV) in the multicenter clinical trial became difficult due to travel ban and social distancing. For multicenter clinical trials, centralized data monitoring is an efficient and cost-effective method of data monitoring. Centralized data monitoring reduces the risk of COVID-19 infections and provides additional capabilities compared to on-site monitoring. The key steps for on-site monitoring include identifying key risk factors and thresholds for the risk factors, developing a monitoring plan, following up the risk factors, and providing a management plan to mitigate the risk.

Methods: For analysis purposes, we simulated data similar to our clinical trial data. We classified the data monitoring process into two groups, such as the Supervised analysis process, to follow each patient remotely by creating a dashboard and an Unsupervised analysis process to identify data discrepancy, data error, or data fraud. We conducted several risk-based statistical analysis techniques to avoid on-site source data verification to reduce time and cost, followed up with each patient remotely to maintain social distancing, and created a centralized data monitoring dashboard to ensure patient safety and maintain the data quality.

Conclusion: Data monitoring in clinical trials is a mandatory process. A risk-based centralized data review process is cost-effective and helpful to ignore on-site data monitoring at the time of the pandemic. We summarized how different statistical methods could be implemented and explained in SAS to identify various data error or fabrication issues in multicenter clinical trials.

1. Introduction

Coronavirus disease 2019 (COVID-19) has affected all kinds of public health operations, including clinical trials. The World Health Organization (WHO) declared COVID-19 as a public health emergency and proclaimed it a pandemic. In clinical trials, it is vital to allocate the medical resources and treatments among the subjects fairly. Due to COVID-19, patients hesitated to visit the clinical sites where they were worried about getting sick or infected by other people. As a result, patients who missed scheduled visits due to COVID-19 related health issues, hospitalizations, self-isolation, quarantine, or lockdown. It became difficult for clinical sites to follow up with the patients and missed clinical information in the current situation. Clinical sites were trying to collect the missing information through alternative methods such as telephone calls. Data monitoring in clinical trials is essential to check data integrity and data quality. There were different types of data errors in clinical trials, and some sort of data errors are more important than others [1,28].

According to International Conference on Harmonization Good Clinical Practice Guideline (ICH GCP) (E6) definition, "Clinical monitoring is the surveillance and regulatory efforts that monitor a participant's safety and efficacy during a trial[14]." Contract research organizations (CRO) monitor clinical trial data by proposing a data monitoring plan and then record, supervise, review, and report the findings of a clinical investigational product. A monitoring plan for an active trial was required to verify the data source through electronic case report verification or reviewing the trial data. Practical and careful data motoring of clinical trials by the sponsor or the contract research organizations played a vital role in protecting human subjects and perform a high-quality study. Sponsors of clinical trials were required to oversee human rights and submit the trial data to the FDA. FDA published several guidelines about how the sponsors should monitor, conduct, and

* Corresponding author.
E-mail address: m.bhuiyan@medpace.com (M.A.N. Bhuiyan).

https://doi.org/10.1016/j.cct.2021.106368
Received 11 January 2021; Received in revised form 22 March 2021; Accepted 23 March 2021
Available online 26 March 2021

progress clinical trials during COVID 19 pandemic[12]. Centralized data monitoring was one of the most efficient and cost-effective methods to ensure human rights, achieve a higher level of data quality, maintain the integrity of the trial, increase patient safety [7,13,28] (Baigent et al., 2015). Due to the COVID-19 pandemic, on-site monitoring became difficult and the clinical sites have restricted on-site monitoring.

Based on US FDA guidelines, risk-based monitoring ensures clinical trial quality by identifying, assessing, monitoring, and reducing the potential risks that may have affected safety and quality [11]. There were three steps in a risk-based monitoring approach; Firstly, identifying the essential data for a study based on the informed consent to inclusion/exclusion screening and mitigating adverse events. Secondly, performing a risk assessment to determine a specific source of risk and the study errors based on those risks. Finally, develop a monitoring plan describing the monitoring method, responsibilities, and the requirements to communicate [9].

We applied the risk-based data monitoring method in a dummy clinical trial data, showed the risk-based monitoring approach through a flow chart and provided several SAS programs, and showed how to detect data discrepancy and interpret the output. We have explained the benefits of the risk-based monitoring approach during the COVID-19 pandemic and mitigate different risks in multicenter clinical trials.

1.1. Data monitoring process

Based on FDA and Clinical Trials Transformation Initiative (CTTI) guidelines, there were different methods to monitor clinical trials depending on trials' focus and methodology [17]. The investigator personnel's regular or periodic on-site visits remained the most favorable data monitoring mechanism in a clinical trial. To maintain FDA data monitoring guidelines, sponsors typically performed source data verification every 4–8 weeks [26]. The coronavirus pandemic hampered the on-site data monitoring process, increased missing data due to participant infection, treatment disruptions, and loss of follow-up. Due to COVID19, the International conference on harmonization (ICH) and the international standards organization (ISO) emphasized centralizing monitoring instead of on-site monitoring, and FDA recommended a risk-based monitoring process [12]. When a sponsor carried out in-person evaluation through sponsor employees at the clinical investigation sites is termed as on-site monitoring. On-site monitoring's primary purpose is to identify the reason for missing data, discrepancies in the case report form (CRF), and the source data and document the information to maintain data integrity and data quality. On-site monitoring helps assess compliance with the study protocol, drug accountability, and an overall idea about the trial's quality. On the other hand, when data evaluation is carried out remotely by the sponsor personnel at a location other than visiting the clinical sites is called centralized data monitoring. Centralized data monitoring provides all the on-site monitoring features and adds some additional advantages [2]. We have shown different steps of a centralized data review process in a flow chart.

1.2. Importance of risk-based data monitoring process

Over the years, risk-based monitoring has grown because it reduced the time-consuming and costly practice of on-site source data verification (SDV) [21]. The fundamental concept of a risk-based central monitoring system was to identify and mitigate risks. A risk assessment in a trial allowed us to identify protocol-specific risks, potential impacts of the risk factors and develop a risk minimization strategy to complete a trial successfully. In clinical trials, risk assessment focused on risks related to subject safety, trial integrity, and data quality. A complete and efficient risk assessment was vital because the regulatory organizations required documents and rationale for a selected central monitoring strategy [23]. Risk-based monitoring reduced the cost of a trial significantly compared to on-site monitoring. Risk-based data monitoring mainly depended on source data verification, a proven resource incentive method with a higher ability to identify and mitigate issues. Because efficient monitoring was essential to maintain patient safety, it was more critical to avoid on-site monitoring to reduce the risk of COVID-19 infection from other patients, clinical staff, or site monitors. There are several reasons why risk-based monitoring was better than on-site monitoring, such as less error, low cost, efficient analysis, cross-site comparison, and timely results [28].

Risk-based centralized monitoring identified the data error through automated reviews [24]. Due to COVID-19, on-site monitoring was not possible and lots of sites restricted on-site monitoring. Centralized monitoring could reduce on-site monitoring and schedule visits to problematic study sites only. We could create a central risk dashboard to statistically and graphically check the data through centralized and risk-based monitoring to determine study outliers and inliers and identify any unusual patterns present in the data. Another advantage of a centralized review system was that it provided a cross-site comparison to assess the site performance and potential fraudulent data identification or miscalibrated data. Risk-based monitoring dashboard made it easier to identify and mitigate any ongoing trial issue too. A dashboard was prepared to monitor the data as the trial went on. For each study, specific risk factors such as vital signs, adverse events, and serious adverse events could be monitored through the dashboard. If a site on the dashboard showed any irregular data or risk factors, the investigator could further investigate on-site data verification to in-depth statistical analysis [22].

2. Methods

We classified risk-based centralized data monitoring into two groups: [1] the supervised monitoring process and [2] the unsupervised monitoring process. We intended to implement a range of data monitoring processes for efficient data monitoring plans and reduce time-consuming and costly on-site source data verification. Our first goal was to follow-up all the subjects in a trial remotely using a patient data dashboard, identify trends and outliers from high-risk sites and perform statistical analysis to check possible data discrepancy. At the time of the COVID-19 pandemic, on-site monitoring was not safe. Patients could also miss scheduled visits and visited the sites during weekends due to travel restrictions or social distancing. We showed that a supervised data monitoring process could identify critical features such as patients' enrollment violations, protocol deviation, adverse events, travel restrictions by region or site, missing visit or dose due to COVID- 19 pandemic, and overall site compliance. We also showed different unsupervised analysis process such as descriptive statistics to identify outliers or influential observations; Pearson correlation analysis between the key variables to check whether values are copied across sites, digit preference to identify fabricated numbers, one way ANOVA to compare different groups that significantly differ from each other, and Chi-square goodness of fit to check a difference in frequency distributions for two or more observations.

3. Results

3.1. Supervised monitoring

To monitor the data collection process and ensure data quality, the supervised analysis could be utilized. Through supervised data monitoring (SDMP), sponsors could monitor protocol compliance and maintained a fair data collection process. For example, following the study visit time point, drug dosing date, inclusion/exclusion criteria, and missing data could be controlled through SDMP. The main idea for a supervised data monitoring process was to identify the key risk factors on different sites of a trial, such as the reason for missing visit or dose, loss of follow-up, incorrect inclusion in the study drug, duration of study drug, or reported adverse events due to pandemic. Key risk factors could

be chosen based on the risk that might impact the study's quality [24]. During the risk assessment process, a tolerance threshold was determined for each factor so that when a specific risk factor has fallen beyond the limit, the sponsors could detect the site, analyze the cause through central monitoring, and adopt necessary action to mitigate the risk [19]. There are several ways to select the cut-off points in the literature, such as ±15% of median, ± Standard deviation of the median, ± median absolute deviation of the median, etc. [19]. The plot in Fig. 2 shows the progress each patient has made in the study. Each vertical line represents the expected visit date, and each point represents the actual visit date, which allowed us to quickly see if patients had their visit within the ±2 day window. The plot is colored by whether or not they received their dose for that visit (legend is located on the bottom). Lastly, the arrow-shaped points indicate today's date, which shows how far each patient is from the previous visit and how soon they are expected to make their next visit. (See Fig. 1.)

3.2. Unsupervised monitoring

In the unsupervised analysis process, the different statistical analysis method was used to identify outliers, inliers, or any specific data pattern without any preconception. In the unsupervised analysis process, there was no predefined threshold or limit. The univariate or multivariate statistical analysis method was used to test the differences in distribution and identify the risk factors [4]. The unsupervised analysis process's key advantages include but are not limited to: efficient tool for multicenter studies and extensive data monitoring and data mining, easy to monitor outliers or extreme values and compared between sites, less time consuming and less costly [28].

3.2.1. Descriptive statistics and outlier detection

Descriptive statistics and different data visualization techniques were used to identify outliers or influential observations. More than 1.5* interquartile range from 75 percentile were considered outliers in a box plot value. Fig. 3 showed a box plot of Hearts rate as one of the vital sign variables and plotted by state and study site to monitor the data. This box plot provided quick insights on where the outliers were and how skewed the results could be for each parameter for single or multisite studies. Multiple vital sign parameters could be monitored and compared simultaneously which reduced a lot of risk factors and ensured efficient study results. Other than a box plot, using a histogram, univariate distribution table, extreme value, distribution plots for subgroups, multivariate scatter plots, or influence statistics could also be helpful statistical methods to identify outliers. Fig. 3(a) showed the box plot of a vital sign parameter by state, and Fig. 3(b) showed the heatmap of the correlation matrix of five vital sign parameters. Based on Fig. 3(a), we saw some data issues for different sites for the state "C" and "P". After checking the frequency distribution, we saw lots of missing data on those states. Fig. 3(b) indicated the correlation matrix of the vital sign parameters for one site (site 110 for state M) and showed the respiratory rate negatively associated with heart rate and systolic blood pressure, which aligns with the other published papers. So, on-site location '110' might have a data fraud issue and we need to check the missing data issues for the state "C" and "P". We could also use cluster heatmap to identify similar sites or subjects with similar data sets for a large number of sites and patients [18].

3.2.2. One-way analysis of variance

One-way ANOVA test was used to test the mean difference between two or more independent and normally distributed observations with equal variance and assumed the residuals follow a normal distribution. The normality assumptions of the residuals could be tested using histograms or Q-Q plots, influential data points were identified through Cook's distance, and equality of variance could be tested using the test of homogeneity (such as folded *t*-test). Post hoc statistical analysis helped to compare different groups that significantly differ from each other.

3.2.3. Chi-square goodness of fit test

The Chi-square goodness of fit test was used to test a difference in frequency distributions for two or more observations. The sum of the squared difference between the observed and the expected values followed a chi-square distribution with a degree of freedom which corresponds to the number of observations considered during the Chi-square calculation. Like *Z*-score, Chi-square distribution was also used to identify a discrepancy in the data. Expected values were based on real data and calculated data. They were not supposed to lie too far or too close to the expected value. The chi-square statistic *p*-value could identify data fabrication or data error if the data too large or too small. There were two methods, such as digit frequency analysis and inlier analysis, that utilize chi-square statistics to identify whether the data lied too far or too close to the expected value or not.

3.2.4. Digit frequency analysis

Human nature favored specific digits during data fabrication [3].

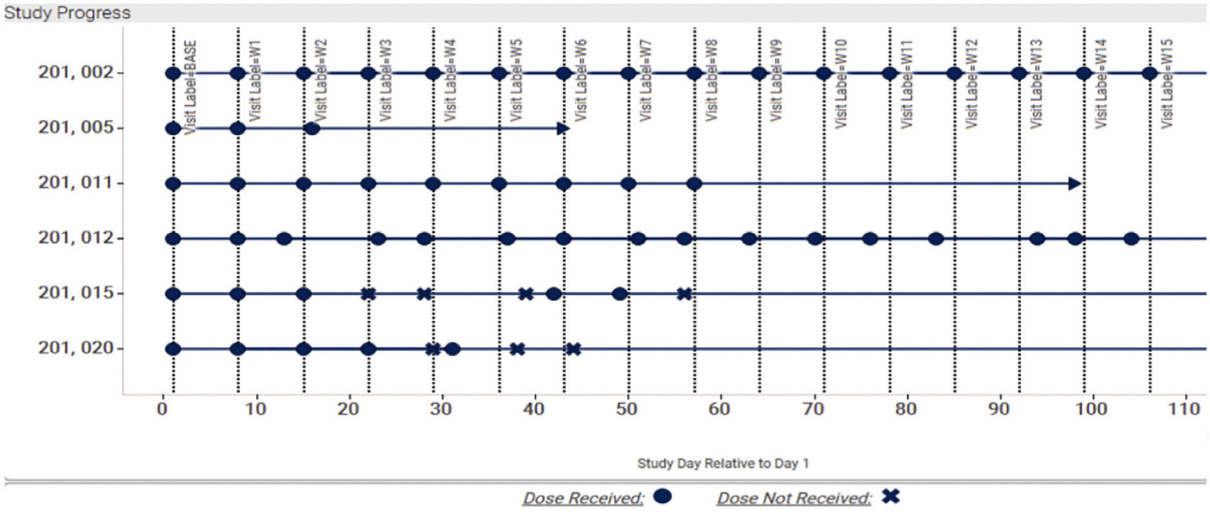

Fig. 2. Ongoing Monitor study visits through a supervised monitoring dashboard. The "X" and blue circle indicate if the patient missed the dose for a specific visit or not. Here, "X" indicates if the patient missed a dose and a blue circle if they received their dose. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

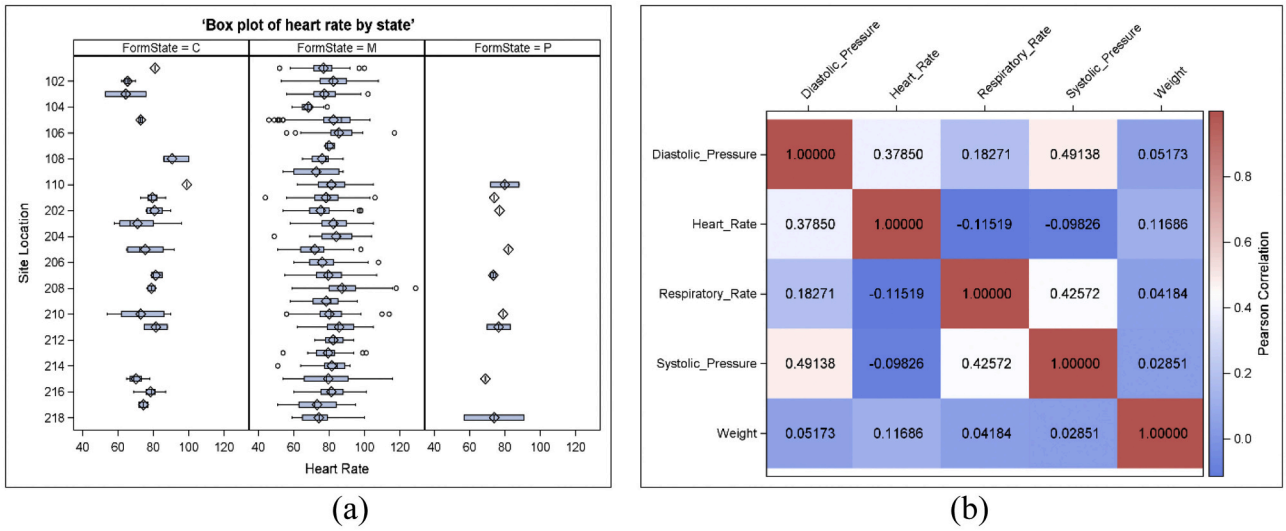

Fig. 3. (a) Box plot of Vital sign parameters by state, (b) Pearson correlation matrix plot for the vital signs.

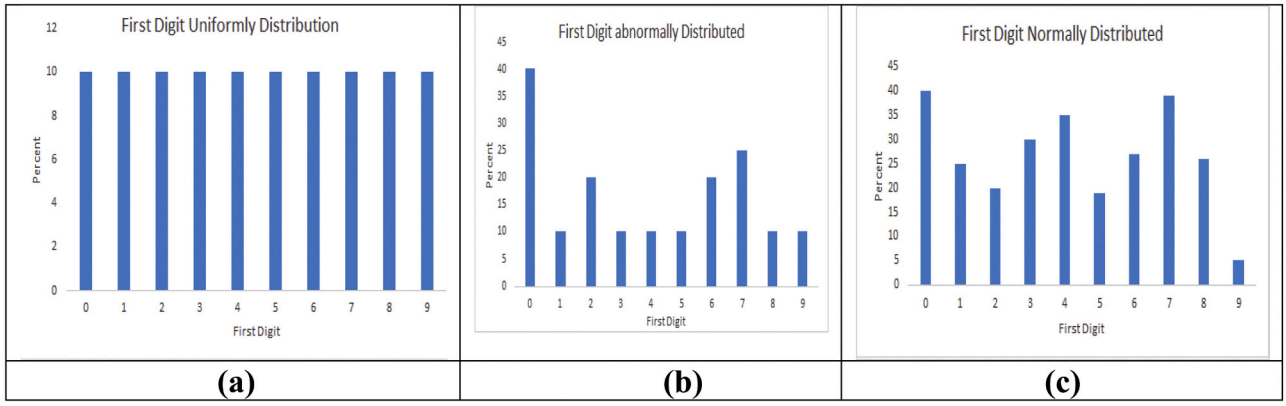

Fig. 4. Examples of first-digit distribution with Chi-square *p*-value. (a), The first digit followed a perfect uniform distribution with a *p*-value of 1.00 (b), where the first digit followed an abnormal distribution with *p*-value 0.0001, and (c), where the first digit followed the normal distribution with a *p*-value of 0.001.

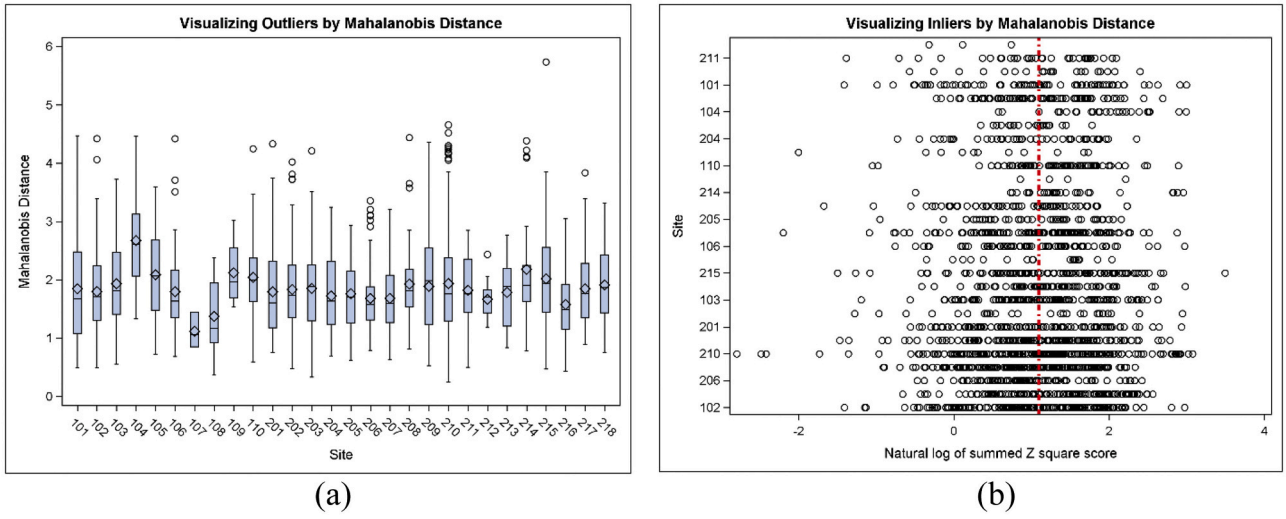

Fig. 5. Mahalanobis distance by study sites. (a) Large distance identified outliers, (b) Small distance determined inliers.

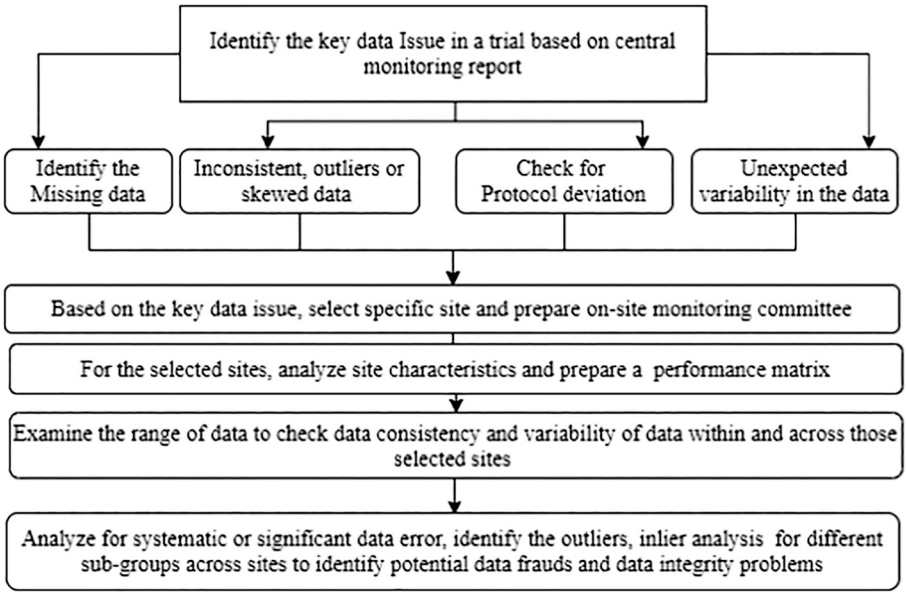

Fig. 1. Flow chart for centralized data review process.

Digit frequency analysis such as First digit analysis [5,27], Last digit analysis [3,8], or other digit analysis method could be utilized to detect an alleged data fraud. We used terminal data analysis using the chi-square goodness of fit formula, where the observed value was the number of times the digit between 0 and 9 appeared in all the sites as the first digit. Each digit's expected values correspond to the number of measurements considered divided by 10 because the equal frequency is assumed for each digit. The degree of freedom considered for terminal digit analysis is the total number of digits −1, which was 9. We took the Respiratory variable in our simulated data and identified the variable's first digit for each site. We used Benford's law [5,27] in our simulated data and showed three different digit analysis scenarios. When a terminal digit, either first or last, appeared more frequently than the other, it may be because of rounding to the nearest digit or data fabrication. Terminal digit analysis could help monitor the data quality and avoid data fraud or data errors.

3.2.5. *Inlier analysis*

In real life, data was expected to vary to a certain extent. The data which lied in the interior of a statistical distribution due to error is called inlier. Inliers were challenging to distinguish from good data points and required additional concern to identify and correct. Mahalanobis distance was a multidimensional risk assessment method that combined a multidimensional risk score [6]. Mahalanobis distance was reduced to the Euclidian distance when normalized by the variance [29]. When the distance was computed for a univariate standard normal distribution, the distance was reduced to the (positive) z-score. Therefore, Mahalanobis distance was the most flexible and optimal method to combine dimensions and identify risk factors. Mahalanobis distance was a measure of distance and the mean away from zero (and not towards zero) was considered relevant. Mahalanobis distance helped identify inliers, indicating data fraud because it was less likely that a subject was an average for a broad set of variables [10]. We used our dummy dataset to create a Mahalanobis distance. Fig. 4 showed a boxplot of distance measured for each site compared to the multivariate distribution centroid. The distance measure was computed from 10 possible variables. Fig. 5 represented Mahalanobis distance by the site to identify outliers (large distance) or inliers (small distance). This plot uncovered possible data errors and depicted critical differences in the study population across sites. In the figure, large variability indicated a diverse study population or location, and low variability indicated a homogeneous population and identified any locations that required further investigation. Fig. 5(a) showed outliers based on Mahalanobis distance based on three vital sign parameters (diastolic blood pressure, systolic blood pressure, Heart rate), and we could see the outliers in site 215 but it didn't show the inliers information. In Fig. 5(b), the inlier based on Mahalanobis distance is shown. The vertical red line indicated the natural log of the degrees of freedom. Based on the distance, the smaller values indicated an inlier. We could see sites 210, 107, 205, and 115 showed inliers and required further investigations.

4. Discussion

Over the decade, the number of clinical trials and their methodological complexities evolved adequately, and clinical trials became genuinely global involving hundreds of sites from several countries. These trials' critical challenges were to confirm effective data monitoring, patient safety, and right and good data quality. The bottlenecks were vast geographic dispersion of the trials, many investigational variables, site support, different treatment, and inequality in standard of care. The traditional monitoring techniques of sponsors on-site visit was becoming expensive, time-consuming and burdensome. This on-site monitoring system even became impossible during the pandemic due to travel restrictions and social distancing. Risk-based monitoring identified the critical risk factors and helped to protect any scientific misconduct or data fraud. Intended or regular data fraud was not usual but may have significantly affected a trial's integrity and jeopardize the life of other patients [16]. One of the standard data fabricated techniques was copying and pasting the existing data within or across study subjects. In such cases, a specific data pattern was visible in the study data and some values occurred more often than others. Each observation's frequency distribution could detect this kind of data fabrication and identified whether a value was copied to different subject charts or split into two groups [4,10].

Previously, several data monitoring ideas have been proposed, such as calendar check to find data error in dates or to identify any trend during the weekend or holidays [1,4]. In other studies, patient-level data monitoring was suggested by examining the date of randomization, vital sign test results, or error in follow-up appointment dates [15]. Kirkwood et al. suggested data monitoring methods using the R programming language. R was a free software programming language and a software environment for statistical computing and graphics. But R was not

widely used in Contract Research Organization or Pharmaceutical companies because of distrust of freeware. Moreover, for big data, R software became slow because it performed operations with everything in memory. In contrast, SAS could better handle big data and, more importantly, FDA used and preferred SAS over any other programming language. Timmermans et al. [25] used SMART™ software to monitor data quality and consistency in the Stomach Cancer Adjuvant Multi-Institutional Trial groups. They created a Data Inconsistency Score (DIS) based on the *p*-value associated with a triplet (center x variable x test). This was an expensive way of data monitoring with a limited scope of identifying correct data discrepancies. We have used SAS to perform the risk-based analysis, which is a more prominent language in FDA and pharmaceutical companies due to ease of learning, robust statistical analysis platform, and interaction with mu; multiple host systems. We have shown how risk-based statistical monitoring could ensure the clinical trial's quality by finding, following, and decreasing the risk that could affect the study's quality or safety. This method was less costly and less time-consuming. We have also shown a flow chart to show the different steps of a risk-based centralized data review process to better facilitate the data monitoring process and develop an appropriate data monitoring plan.

COVID-19 pandemic raised different intercurrent events in a clinical trial. It is essential to assess several vital points during an ongoing study such as patient recruitment and treatment availability, data quality and data integrity, and the trial's overall feasibility. The sponsors needed to take the necessary actions to reduce missing data due to COVID-19. Risk-based centralizing monitoring's main objective is to ensure patient safety and maintain data quality and data integrity most efficiently and cost-effectively. Risk-based centralized monitoring played an essential role during COVID-19 by reducing the risk of getting infected. During the trial, a patient dashboard allowed to follow up a patient remotely. For multicenter studies, the sponsor can reduce the time and cost by monitoring and identifying the sites facing different issues related to COVID-19, such as patients who are missing schedule visits, study drugs, or facing adverse effects, and take necessary steps to mitigate the risk. It is more helpful and easy to monitor patient enrollments and data discrepancy for multisite and multi-location studies following our supervised data monitoring process. Several data errors can occur due to missing scheduled visits or missing dose, and this type of error can be reduced through supervised monitoring. Supervised data visualization at both patient and site-level will help to perform further unsupervised statistical analysis on the selected cites. Finally, a risk-based monitoring system can reduce one-fourth of the trial cost [20] and increase trial efficiency. The supervised analysis process will identify the critical risk factors, and later unsupervised analysis can be used to identify data discrepancy.

Declaration of competing interest

The corresponding author states that, on behalf of all authors, there is no conflict of interest.

Appendix A. Appendix

```
/*Creating Box Plot by different Subgroup S.A.S. code*/.
PROC SGPANEL DATA = DATA_NAME;
PANEL BY XX (SUBGROUP VARIABLE NAME);
VBOX YY (VARIABLE NAME);
RUN;
/*Creating Multivariate Scatter Plot by different Subgroup S.A.S. code*/
PROC SGPANEL DATA = DATA_NAME;
PANEL BY XX (XX, SUBGROUP VARIABLE NAME);
SCATTER Y=VARIABLE_NAME X = VARIABLE_NAME;
RUN;
/* Finding Influential statistics in SAS */
ODS OUTPUT LSMEANS = LSMEANSOUT;
PROC GLM DATA = DATA_NAME;
CLASS CLASS_VARIABLE;
MODEL RESPONSE = CLASS_VARIALE OTHER_VARIABLE;
LSMEANS CLASS_VARIABLE/PDIFF STDERR;
OUTPUT OUT = COOKSD (WHERE = (COOKSD≥ 4/&NOBS)) COOKD=COOKSD;
QUIT;
ODS OUTPUT CLOSE;
/* Finding univariate Mahalanobis distance in SAS */
%MACRO MAHALANOBIS_DIS_UN(DATA=, ID=, VAR=);
PROC PRINCOMP DATA = &DATA1 STD OUT = DATA1 OUTSTAT = OUTSTAT;
VAR &VAR;
RUN;
DATA DATA2;
SET DATA1;
&VAR.D=SQRT (USS(OF PRIN:));
RUN;
PROC SORT DATA = DATA1; BY &ID; RUN;
PROC SORT DATA = DATA2; BY &ID; RUN;
DATA MAHALANOBIS;
MERGE DATA1 DATA2;
BY &ID;
RUN;
%MEND;
% MAHALANOBIS_DIS_UN (DATA = DATA_NAME, ID = ID_VARIABLE, VAR = VARIABLE_NAME);
/* Finding multivariate Mahalanobis distance in SAS */
%MACRO MAHALANOBIS_DIS_MUL(DATA=, ID=, VAR1=, VAR2=, VAR3=);
PROC PRINCOMP DATA = &DATA1 STD OUT = DATA1 OUTSTAT = OUTSTAT;
VAR &VAR1 &VAR2 &VAR3;
RUN;
DATA DATA2;
SET DATA1;
&VAR &VAR2 &VAR3.D=SQRT (USS(OF PRIN:));
DROP PRIN;
RUN;
PROC SORT DATA = DATA1; BY &ID; RUN;
PROC SORT DATA = DATA2; BY &ID; RUN;
DATA MAHALANOBIS;
MERGE DATA1 DATA2;
BY &ID;
RUN;
%MEND;
% MAHALANOBIS_DIS_MUL (DATA = DATA_NAME, ID = ID_VARIABLE, VAR1 = VARIABLE1, VAR2 = VARIABLE2, VAR3 = VARIABLE3);
/***** Terminal Digit analysis*****/
ODS HTML STYLE = STATISTICAL;
ODS GRAPHICS ON;
DATA D;
SET D1;
DIG_LAST = mod(VAR_NAME,10); * select the last digit
DIG_FIRST = SUSTR(PUT(VAR_NAME,best3.),1,2); * select the first digit
RUN;
/**********Last Digit**********/
PROC FREQ DATA = D;
WHERE SITE in ('SITE_NAME' 'SITE_NAME');
TABLES SITE*DIG_LAST /chisq plots = freqplot;
run;
ODS GRAPHICS OFF;
ODS HTML CLOSE;
```

```
/**********FIRST Digit**********/
PROC FREQ DATA = D;
WHERE SITE in ('SITE_NAME' 'SITE_NAME');
TABLES SITE*DIG_FIRST/chisq plots = freqplot;
run;
ODS GRAPHICS OFF;
ODS HTML CLOSE;
```

References

[1] 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, Clinical Trials 5 (2008) 49–55.

[2] J.M. Bakobaki, M. Rauchenberger, N. Joffe, S. McCormack, S. Stenning, S. Meredith, The potential for central monitoring techniques to replace on-site monitoring: findings from an international multicenter clinical trial, Clinical Trials 9 (2) (2012) 257–264.

[3] A. Beauregard, V. Tantsyura, F. Labrie, The basics of clinical trial centralized monitoring, Appl. Clin. Trials 27 (11) (2018).

[4] M. Buyse, S.L. George, S. Evans, N.L. Geller, J. Ranstam, B. Scherrer, T. Colton, The role of biostatistics in the prevention, detection, and treatment of fraud in clinical trials, Stat. Med. 18 (24) (1999) 3435–3451.

[5] J.C. Collins, Using excel and ''Benford's law to detect fraud: learn the formulas, functions, and techniques that enable efficient benford analysis of data sets, J. Account. 223 (4) (2017) 44.

[6] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis Distance, Chemometrics, and Intelligent Laboratory Systems, 2000.

[7] Dimitris K. Agrafiotis, Victor S. Lobanov, Michael A. Farnum, Eric Yang, Joseph Ciervo, Michael Walega, Adam Baumgart, Aaron J. Mackey, Risk-based monitoring of clinical trials: an integrative approach, Clin. Ther. 40 (7) (2018) 1204–1212.

[8] S. Dlugosz, U. Müller-Funk, The value of the last digit: statistical fraud detection with digit analysis, ADAC 3 (3) (2009) 281.

[9] European Medicines Agency, Reflection paper on risk-based quality management in clinical trials, Compliance Insp 44 (2013) 1–15.

[10] S. Evans, Statistical aspects of the detection of fraud, in: Fraud and Misconduct in Medical Research, 3rd ed, BMJ Publishing Group, London, 2001, pp. 186–204.

[11] Food and Drug Administration, Guidance for Industry: Oversight of Clinical Investigations—A Risk-Based Approach to Monitoring, FDA, Silver Spring, MD, 2013.

[12] Food and Drug Administration, Guidance on the Conduct of Clinical Trials of Medical Products During COVID-19 Pandemic Guidance for Industry, Investigators, and Institutional Review Boards, Available at, www.fda.gov, 2020.

[13] S.L. George, M. Buyse, Data fraud in clinical trials, Clin. Investig. (Lond). 5 (2) (2015) 161–173.

[14] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, Integrated Addendum to ICH E6 (R1): Guideline for Good Clinical Practice E6 (R2), 2016.

[15] A.A. Kirkwood, T. Cox, A. Hackshaw, Application of methods for central statistical monitoring in clinical trials, Clin. Trials 10 (5) (2013) 783–806.

[16] S. Lock, Fraud and misconduct in medical research, Med. Leg. J. 64 (4) (1996) 139–152.

[17] B.W. Morrison, C.J. Cochran, J.G. White, J. Harley, C.F. Kleppinger, A. Liu, J. D. Neaton, Monitoring the quality of conduct of clinical trials: a survey of current practices, Clin. Trials 8 (3) (2011) 342–349.

[18] V. Novak, P. Novak, J. de Champlain, A.R. Le Blanc, R. Martin, R. Nadeau, Influence of respiration on heart rate and blood pressure fluctuations, J. Appl. Physiol. 74 (2) (1993) 617–626.

[19] K. Oba, Statistical challenges for central monitoring in clinical trials: a review, Int. J. Clin. Oncol. 21 (1) (2016) 28–37.

[20] A. Sertkaya, H.H. Wong, A. Jessup, T. Beleche, Key cost drivers of pharmaceutical clinical trials in the United States, Clin. Trials 13 (2) (2016) 117–126.

[21] L.B. Sullivan, The Current Status of Risk-Based Monitoring, 2015.

[22] V. Tantsyura, I. Grimes, J. Mitchel, K. Fendt, S. Sirichenko, J. Waters, B. Tardiff, Risk-based source data verification approaches pros and cons, Drug Inf. J.: D.I.J./ Drug Inf. Assoc. 44 (6) (2010) 745–756.

[23] R. Temple, Policy developments in regulatory approval, Stat. Med. 21 (19) (2002) 2939–2948.

[24] TransCelerate BioPharma Inc., Position Paper: Risk-Based Monitoring Methodology, 2013.

[25] C. Timmermans, E. Doffagne, D. Venet, L. Desmet, C. Legrand, T. Burzykowski, M. Buyse, Statistical monitoring of data quality and consistency in the Stomach cancer Adjuvant Multi-institutional Trial Group Trial, Gastric Cancer 19 (1) (2016) 24–30.

[26] R.W. Usher, PhRMA BioResearch Monitoring Committee perspective on acceptable approaches for clinical trial monitoring, Drug Inf. J. 44 (4) (2010) 477–483.

[27] D.T. Varma, D.A. Khan, Fraud detection in supply chain using benford distribution, Int. J. Res. Manag. 5 (2) (2012).

[28] D. Venet, E. Doffagne, T. Burzykowski, F. Beckers, Y. Tellier, E. Genevois-Marlin, et al., A statistical approach to central monitoring of data quality in clinical trials, Clin Trials. 9 (2012) 705–7013.

[29] D.R. Wilson, T.R. Martinez, Improved heterogeneous distance functions, J. Artif. Intell. Res. 6 (1997) 1–34.