

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Food and Drug Administration Papers

U.S. Department of Health and Human Services

2019

Comparing methods for clinical investigator site inspection selection: a comparison of site selection methods of investigators in clinical trials

Nicholas Hein

University of Nebraska Medical Center, Omaha

Elena Rantou

U.S. Food and Drug Administration, Silver Spring, Elena.Rantou@fda.hhs.gov

Paul Schuette

U.S. Food and Drug Administration, Silver Spring

Follow this and additional works at: <https://digitalcommons.unl.edu/usfda>



Part of the [Dietetics and Clinical Nutrition Commons](#), [Health and Medical Administration Commons](#), [Health Services Administration Commons](#), [Pharmaceutical Preparations Commons](#), and the [Pharmacy Administration, Policy and Regulation Commons](#)

Hein, Nicholas; Rantou, Elena; and Schuette, Paul, "Comparing methods for clinical investigator site inspection selection: a comparison of site selection methods of investigators in clinical trials" (2019). *Food and Drug Administration Papers*. 48.
<https://digitalcommons.unl.edu/usfda/48>

This Article is brought to you for free and open access by the U.S. Department of Health and Human Services at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Food and Drug Administration Papers by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.



Comparing methods for clinical investigator site inspection selection: a comparison of site selection methods of investigators in clinical trials

Nicholas Hein^a, Elena Rantou^b, and Paul Schuette^b

^aDepartment of Biostatistics, University of Nebraska Medical Center, Omaha, NE, USA; ^bOffice of Biostatistics/Office of Translational Sciences/Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA

ABSTRACT

Background During the past two decades, the number and complexity of clinical trials have risen dramatically increasing the difficulty of choosing sites for inspection. FDA's resources are limited and so sites should be chosen with care.

Purpose To determine if data mining techniques and/or unsupervised statistical monitoring can assist with the process of identifying potential clinical sites for inspection.

Methods Five summary-level clinical site datasets from four new drug applications (NDA) and one biologics license application (BLA), where the FDA had performed or had planned site inspections, were used. The number of sites inspected and the results of the inspections were blinded to the researchers. Five supervised learning models from the previous two years (2016–2017) of an on-going research project were used to predict site inspections results, i.e., No Action Indicated (NAI), Voluntary Action Indicated (VAI), or Official Action Indicated (OAI). Statistical Monitoring Applied to Research Trials (SMARTTM) software for unsupervised statistical monitoring software developed by CluePoints (Mont-Saint-Guibert, Belgium) was utilized to identify atypical centers (via a *p*-value approach) within a study. Finally, Clinical Investigator Site Selection Tool (CISST), developed by the Center for Drug Evaluation and Research (CDER), was used to calculate the total risk of each site thereby providing a framework for site selection. The agreement between the predictions of these methods was compared. The overall accuracy and sensitivity of the methods were graphically compared.

Results Spearman's rank order correlation was used to examine the agreement between the SMARTTM analysis (CluePoints' software) and the CISSTTM analysis. The average aggregated correlation between the *p*-values (SMARTTM) and total risk scores (CISST) for all five studies was 0.21, and range from –0.41 to 0.50. The Random Forest models for 2016 and 2017 showed the highest aggregated mean agreement (65.1%) amongst outcomes (NAI, VAI, OAI) for the three available studies. While there does not appear to be a single most accurate approach, the performance of methods under certain circumstances is discussed later in this paper.

Limitations Classifier models based on data mining techniques require historical data (i.e., training data) to develop the model. There is a possibility that sites in the five-summary level datasets were included in the training datasets for the models from the previous year's research which could result in spurious confirmation of predictive ability. Additionally, the CISST was utilized in three of the five site selection processes, possibly biasing the data.

Conclusion The agreement between methods was lower than expected and no single method emerged as the most accurate.



ARTICLE HISTORY

Received 1 October 2018

Accepted 18 June 2019

KEYWORDS

Site inspection; data mining; unsupervised statistical monitoring; risk assessment; *p*-values

CONTACT Elena Rantou  Elena.Rantou@fda.hhs.gov  Office of Biostatistics/Office of Translational Sciences/Center for Drug Evaluation and Research, U.S. Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lbps.

1. Background

The U.S. Food and Drug Administration (FDA) is responsible for making decisions about new drug applications (NDAs) and biologics license applications (BLAs). The reliability and integrity of clinical trial data is crucial to such marketing approval decisions. The Office of Scientific Investigations (OSI), in the Office of Compliance, Center for Drug Evaluation and Research (CDER), collaborates with other FDA offices to verify data quality and data integrity submitted to CDER in support of the NDAs and BLAs. As part of the review process, CDER reviewers may request Office of Regulatory Affairs (ORA) investigators conduct on-site inspections at selected clinical investigator sites. On-site inspections help ensure that the safety and efficacy of experimental treatments are accurately assessed. However, FDA's resources are limited. In addition to FDA inspections, sponsors are expected to monitor clinical trial sites and are responsible for the integrity and quality of submitted data (see [U.S. FDA](#), Section 5.18) .

Due to limited FDA resources, targeted selection of sites which are potentially most problematic is desirable. In 2010, CDER began a pilot of its Clinical Investigator Site Selection Tool (CISST), a risk-based model to identify clinical investigator sites for inspection. More recently, CDER (with the support of Oak Ridge Institute for Science and Education (ORISE) interns) has explored whether supervised data mining methods can be used to predict the outcomes of a site-inspection. Currently, through a Cooperative Research and Development Agreement (CRADA), CDER is exploring the use of unsupervised statistical monitoring software (Venet and Doffagne 2016; Trotta et al. 2019) developed by CluePoints to identify atypical sites, which may be indicative of data quality and integrity issues. The purpose of this study is to determine whether the various tools (i.e., CISST, supervised data mining methods, and the CluePoints software) broadly agree in their respective findings and to determine which tool, if any, is the most accurate.

2. Data

We used five summary-level clinical site datasets from four NDAs and one BLA where the FDA had performed on-site inspections. The data are based on the `clinsite.xpt` data set described in (U.S. Food and Drug Administration 2018a) Applications were labeled A, B, C, D, and E (in no particular order) to ensure confidentiality of the data. The authors were blinded to the number of inspections performed and the results of the inspections. The five applications were submitted between 2013 and 2018. Sites with zero enrollment were excluded from the analyses. The average number of sites per application was 304 with a median enrollment of 13 subjects across all five applications. The smallest application had 29 sites with a total patient enrollment of approximately 135 individuals, while the largest application enrolled approximately 15,000 patients across nearly 700 sites.

2.1. Outcomes

The results of an on-site inspection are reported as No Action Indicated (NAI), Voluntary Action Indicated (VAI), or Official Action Indicated (OAI).

2.1.1. No action indicated (NAI)

An NAI classification means no objectionable conditions or practices were found during the inspection.

2.1.2. Voluntary action indicated (VAI)

A VAI classification indicates that objectionable conditions or practices were found but the agency is not prepared to take or recommend any administrative or regulatory action.

2.1.3. Official Action Indicated (OAI)

An OAI classification occurs when objectionable conditions or practices were found and regulatory and/or administrative actions will be recommended. (see U.S. FDA, 2018c)

Empirical evidence suggests that an occurrence of OAI is a rare event with PDUFA/BSUFA inspections; approximately 1% of clinical sites being classified as OAI (Jha et al. 2017; Tang et al. 2016).

2.2. Missing data

Summary-level clinical site data (clinsite.xpt) comprised of demographic information and Key Risk Indicators (KRI) of safety and efficacy were used for the analysis tools. For the studies and KRI's of interest, rates of missing data varied from 0% for some variables of interest, to as much as 100% for financial disclosure in some studies.

According to Schafer and Graham (2002), statisticians may classify missing data as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Data are classified as MCAR if the probability of a particular value missing is independent of both the observed and unobserved data (Schafer and Graham 2002). If the probability of a particular value missing depends only on the observed data, then the missing data are classified as MAR (Schafer and Graham 2002). Similarly, Schafer and Graham state that if the probability of a value missing depends on both the observed and unobserved data, the missing data are said to be MNAR.

Data with missing values can be analyzed using complete case analysis (rows with missing data are excluded from analysis) or using imputed data (replacing missing data with a substituted value). Only in the case of MCAR does complete case analysis provide unbiased estimates (Schafer and Graham 2002). Analyses performed using imputed data can produce unbiased estimates in all three cases; however, MNAR data assumes missingness depends on the unobserved data and requires knowledge of the underlying mechanism of the missingness (Schafer and Graham 2002). Using logistic regression, missing data were found to depend on observed data (results not shown). Additionally, since MNAR is an untestable hypothesis, therefore, we assumed data were MAR.

3. Methods

Five supervised learning models from the 2016 and 2017 phases of the research project to predict site inspection outcomes were employed. The two models from 2016 classified sites as NAI, VAI, or OAI (Tang et al. 2016); while the three models from 2017 classified sites as NAI or VAI/OAI (Jha et al. 2017). We used CluePoints' Statistical Monitoring Applied to Research Trials (SMARTTM) software (Venet and Doffagne 2016) to identify atypical sites via *p*-values (see Section 3.3). Lastly, the FDA's CISST was used to calculate the total risk of each site, thereby providing a framework for site selection. See Table 1 for a summary of methods.

The five data mining models from 2016 and 2017 require complete cases to make a prediction. However, SMARTTM and CISST do not require complete cases for their respective analyses. Therefore, we only imputed missing data when performing analyses using the models from 2016 and 2017.

Table 1. Summary of methods used to predict site inspection outcomes.

	SMART TM	CISST	Data mining	
			2016	2017
Description	Detects outliers using distributional assumptions about the data.	Expert opinions used to develop a risk-based model.	Historical data used to train classification models for prediction.	Historical data used to train classification models for prediction.
Predictions	Uses <i>p</i> -value to identify atypical sites.	Assigns risk score to each site.	NAI, VAI, or OAI.	NAI or VAI/OAI

Missing data were imputed using the Random Forest imputation method (Liaw and Wiener 2002) when we performed analyses using the 2016 models. For the 2017 models, the Random Forest imputation method (Liaw and Wiener 2002) was first used to impute missing instances of financial disclosure. We then used the Expectation-Maximization (EM) algorithm (Honaker et al. 2011) to impute the remaining instances of missing values in the variables site efficacy, treatment efficacy, and time since last inspection. The EM algorithm (Honaker et al. 2011) was unable to handle the substantial amount of missingness in the financial disclosure variable. Additionally, we tried to align the imputation methods we used with the imputation methods used in Tang et al. (2016) and Jha et al. (2017).

We examined the percent agreement between the two 2016 and the three 2017 data mining models. To align with predictions from the 2017 models, we re-classified predictions from the 2016 models as NAI or VAI/OAI. We then calculated percent agreement as the number of concordant predictions divided by the total number of predictions. We compared the results of the SMARTTM (Venet and Doffagne 2016) analyses and CISST analyses using Spearman’s rank order correlation.

To examine agreement between all methods and the official outcomes, we performed a visual examination using heat maps (Wickham 2009) of the predictions and the official outcomes. We converted the predictions to a continuous [0,1] scale. CISST risk scores were converted as

$$(score - Min(score)) / (Max(score) - Min(score)).$$

For the data mining methods, we assigned each category a value (Table 2). The SMARTTM *p*-values did not require converting; however, we reversed the *p*-values to align with a larger *p*-value indicating an increased likelihood of atypical sites.

We stratified all analyses by application unless otherwise stated. Additionally, we analyzed sites by combining treatment arms within an application, as well as, by arms separately reporting the worst-case prediction of the arms. SMARTTM analyses used proprietary methods for combining the results of treatment arms within a site. We conducted the analyses utilizing the five supervised learning models using R v3.5.0 (R Core Team 2018) on a Windows 10 PC. We performed the SMARTTM (Venet and Doffagne 2016) analyses using R v3.3.2 (R Core Team 2016) in a Linux high-performance computing environment. Results of the CISST analysis were included in the data as additional columns.

3.1. Data mining methods – 2016

Tang et al. (2016) considered five types of classification models: Boosted Tree, Classification Tree, Combined Binary Random Forest, Ordinal Regression, and Random Forest. Each class of model had external parameters (see Section 3.1.1) that could be tuned resulting in multiple models for each class, e.g., Random Forest, Random Forest + alpha = 0.02231, Random Forest + SMOTE + alpha = 0.093, etc. (Tang et al. 2016). Tang et al. used site inspection data from 2013–2015 containing 3561 sites to train and validate the models using 5-fold cross-validation. We selected the two best performing models for our current study (Table 3). Tang and others defined performance using overall accuracy and sensitivity of predicting OAI. Based on Tang and colleagues’ results, Random Forest + alpha = 0.093 and Boosted Tree + SMOTE + alpha = 0.2231 performed the best. The Random Forest + alpha = 0.093 model will be referenced as Random Forest 2016, and the

Table 2. Numerical assignment for predictions from data mining methods.

2016 Methods		2017 Methods	
Prediction	Assigned Value	Prediction	Assigned Value
NAI	0	NAI	0
VAI	0.5	VAI/OAI	0.8
OAI	1		

Table 3. Accuracy and sensitivity of top performing data mining models from 2016.

Model	Accuracy	Sensitivity
RF + SMOTE	90.54%	81.94%
Boost + SMOTE	91.27%	86.05%
CRF + SMOTE	89.89%	88.37%
RF + alpha = 0.093	89.16%	95.35%
Boost + alpha = 0.093	90.81%	93.02%
CRF + alpha = 0.093	88.88%	97.67%
RF + SMOTE + alpha = 0.2231	89.07%	90.69%
Boost + SMOTE + alpha = 0.2231	89.03%	95.35%
CRF + SMOTE + alpha = 0.2231	88.24%	100%

Note. The following abbreviations are used: RF – Random Forest; Boost – Boosted Tree; CRF – Combined Binary Random Forest; alpha – threshold adjustment; SMOTE – Synthetic Minority Oversampling Technique. Adapted from “Exploring data mining methods for clinical site investigator inspection,” by M. Tang, E. Rantou, and P. Schuette, 2016, unpublished manuscript.

Boosted Tree + SMOTE + alpha = 0.2231 model will be referred to as Boosted Tree 2016 henceforth. Both models had an overall accuracy of 89% and a sensitivity of 95% (Tang et al. 2016).

3.1.1. Imbalance in outcomes

For classification models, the classifier generally takes more information from the majority class. Imbalanced data can cause a problem when using classification methods, e.g., Boosted Tree, Random Forest, Ordinal Regression, etc. For example, suppose the majority class accounts for 99% of the outcomes. Naively classifying all samples as the majority class will result in 99% total accuracy; however, none of the minority class will be correctly predicted. Therefore, adjustments to the minority class are needed to improve the sensitivity of detecting the minority class. This adjustment may lead to slightly lower overall accuracy than naively classifying all samples as the majority class. Two such approaches are Synthetic Minority Oversampling Technique (SMOTE) developed by Chawla et al. (2002) and threshold adjustment.

The external tuning parameters of the models (SMOTE and alpha) were used to compensate for the low percentage of OAI outcomes in the dataset. The 3561 sites had an OAI outcome rate of 1.2% (Tang et al. 2016). SMOTE is a method that generates synthetic samples for a minority class (i.e., OAI) thereby increasing the percentage of OAI outcomes (Chawla et al. 2002). Whereas alpha is a threshold adjustment of the predicted probabilities of the respective model. The final classification of a site is determined by the class that gives the largest probability,

$$f(X_i) = \operatorname{argmax}_{j \in \{NAI, VAI, OAI\}} \alpha_j P(Y_i = j),$$

where $\alpha_{OAI} = 1$ and $\alpha_{NAI} = \alpha_{VAI} = \{0.093, 0.2231, 1\}$.

Thus, alpha is used as a scalar for the probabilities of NAI and VAI thereby increasing the likelihood the $P(Y_i = OAI)$ is the maximum, and hence adjusting for the imbalance in the percentage of outcomes.

3.1.2. Predictors

Scientists from OSI identified 13 KRIs believed to influence the site inspection results, as identified in Bioresarch Monitoring Technical Conformance Guide (US FDA 2018a), and supplemented by internal FDA data. These 13 KRIs were used to predict NAI, VAI, or OAI. Three KRIs had missing

Table 4. Validation error and percent misclassification of 2017 data mining models.

Method	Cross Validation Error	Misclassification
Random Forest	13.5%	14.0%
Boosted Tree	15.9%	14.9%
Boosted Tree with Dropout	16.9%	16.4%

Note. Adapted from “Data mining tool for clinical site investigator inspection,” by C. Jha, E. Rantou, and P. Schuette, 2017, unpublished manuscript.

data (Tang et al. 2016). Tang and colleagues (2016) used Random Forest imputation method to impute the missing data.

3.2. Data mining methods – 2017

Three classification models were considered in 2017 and subsequently chosen for our study, Random Forest, Boosted Tree, and Boosted Tree with Dropout (Jha et al. 2017). Jha et al. (2017) used site inspection data from 2013–2016 with a total of 3035 entries to train and validate the models. OAI outcome prevalence was <1% (Jha et al. 2017). Jha and colleagues accounted for this imbalance by dichotomizing the outcomes, i.e., combining the VAI and OAI outcomes into one category. This dichotomization resulted in 64.2% NAI outcomes and 35.8% VAI/OAI outcomes (Jha et al. 2017). Fifteen KRIs were used to predict the binary outcomes NAI or VAI/OAI. The 15 KRIs included the 13 KRIs from 2016 in addition to treatment efficacy and site-specific efficacy missing indicator. Jha and colleagues found four KRIs contained missing data; the missing data were imputed using the EM algorithm (Honaker et al. 2011).

Jha et al. (2017) used a simple random sample without replacement to split the data into training and validation data using a sampling probability of 0.8 for the training data. Jha and others defined model performance using 5-fold cross-validation error using the training data and percent misclassification using validation data (Table 4).

3.3. Statistical monitoring applied to research trials (SMARTTM)

SMARTTM (Venet and Doffagne 2016) is a statistical software package designed to detect atypical sites in a multicenter study. SMARTTM (Venet and Doffagne 2016) assumes that data coming from the various centers of the study are similar except for some natural variations due to chance or variations due to the design of the study, e.g., multinational study that recruits patients with different ethnicities (Venet et al. 2012). SMARTTM software tests the distribution of data in one center to data in all other centers (Venet et al. 2012). A number of p -values are returned from these tests (Venet et al. 2012). A weighted geometric mean, P , of the p -values for the center is then computed; where the weights reflect the correlation between the tests that were performed. The data inconsistency score (DIS) for a center is then calculated as $DIS = -\log(P)$. We flagged a center as atypical if the site weighted geometric mean P was less than 0.05.

SMARTTM (Venet and Doffagne 2016) was initially designed to analyze patient-level data. Recently, CluePoints modified some of their statistical tests used in the SMARTTM (Venet and Doffagne 2016) software for summary-level data. We examined the five applications of this study using both patient and summary level data. For the summary level data, we tested 10 KRIs for inconsistencies across centers. The results of the SMARTTM (Venet and Doffagne 2016) analyses are from summary-level data.

3.4. Clinical investigator site selection tool (CISST)

CISST is a tool developed by eliciting expert opinion. Twenty-one attributes that are potentially associated with the outcome of on-site inspections were identified through a series of expert interviews across the offices of Biostatistics (OB), Office of Compliance (OC), Office of New Drugs (OND), and Office of Business Informatics (OBI). In the second round of interviews, the identified attributes were ranked in their importance. A risk-based algorithm was then developed. The algorithm transforms the KRIs, applies weights to each risk function, and aggregates each risk factor into a total risk score. We used the total risk score as an indicator of the likelihood of OAI outcome. The higher the risk score, the higher the probability that a site inspection would result in an OAI classification.

4. Results

We examined the agreement between SMARTTM (Venet and Doffagne 2016) analyses using patient-level SDTM data and summary-level (clinsite) data via Spearman's rank order correlation (Figure 1). The average aggregated correlation between the *p*-values of the two analyses for the five applications was 0.29. There was no association between the magnitude of the correlation and the number of patients enrolled in the study.

Spearman's rank order correlation was also used to determine the agreement between the SMARTTM (Venet and Doffagne 2016) analyses (summary-level data) and the risk scores from CISST. We first analyzed the data using the CISST by combining the results of each treatment arm and calculated the total risk score for the combined results. We will refer to this method of analysis as 'combined' analysis. Similarly, we analyzed the data using the CISST by calculating a total risk score for each arm and reporting the maximum total risk score of the arms for the respective site. We refer to this method as 'by arm' analysis. The average aggregated correlation between SMARTTM (Venet and Doffagne 2016) *p*-values and CISST total risk scores across the five applications was -0.21 and -0.19 for combined and by arm analysis, respectively (Figure 2). We expected a negative

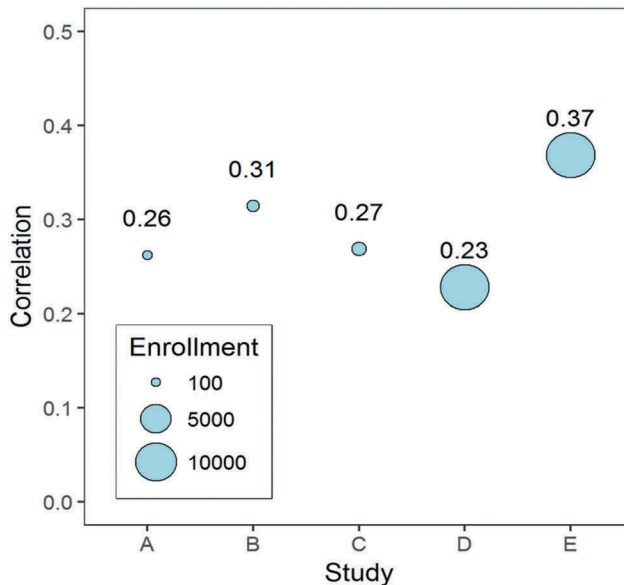


Figure 1. Spearman's correlation between SMARTTM analyses *p*-values.

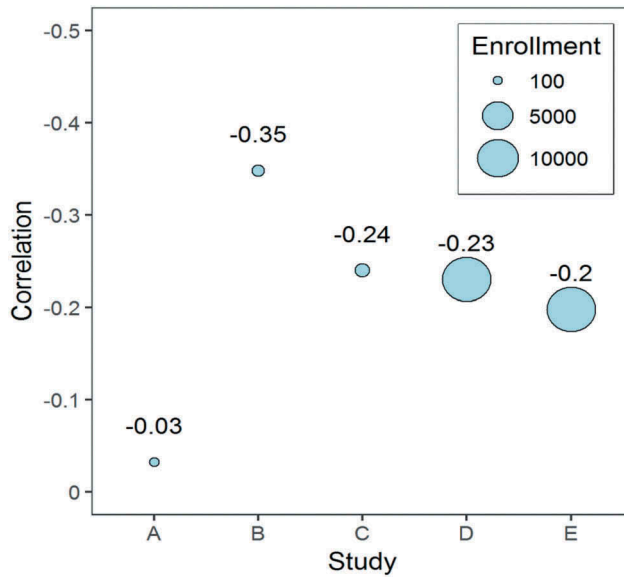


Figure 2. Spearman’s correlation between SMART™ *p*-values (summary-level data) and CISST risk scores.

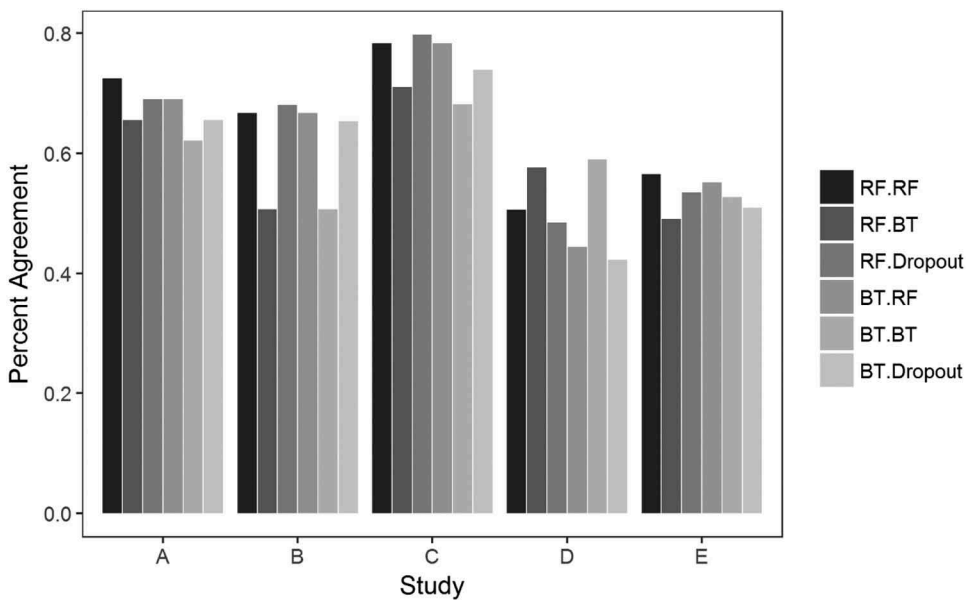


Figure 3. Percent agreement between data mining methods of 2016 and 2017. The legend is designed as follows: [model].[model] where the first [model] is a 2016 model and second [model] is a 2017 model. The following abbreviations are used: RF – random forest; BT – boosted tree; Dropout – boosted tree with dropout.

correlation since a small *p*-value from SMART™ (Venet and Doffagne 2016) indicates an atypical site and a sizable total risk score from CISST suggests a potential problem site.

Next, we examined the percent agreement between the six combinations of models from 2016 and 2017 (Figure 3). The Random Forest models from 2016 and 2017 (combined analysis) showed the

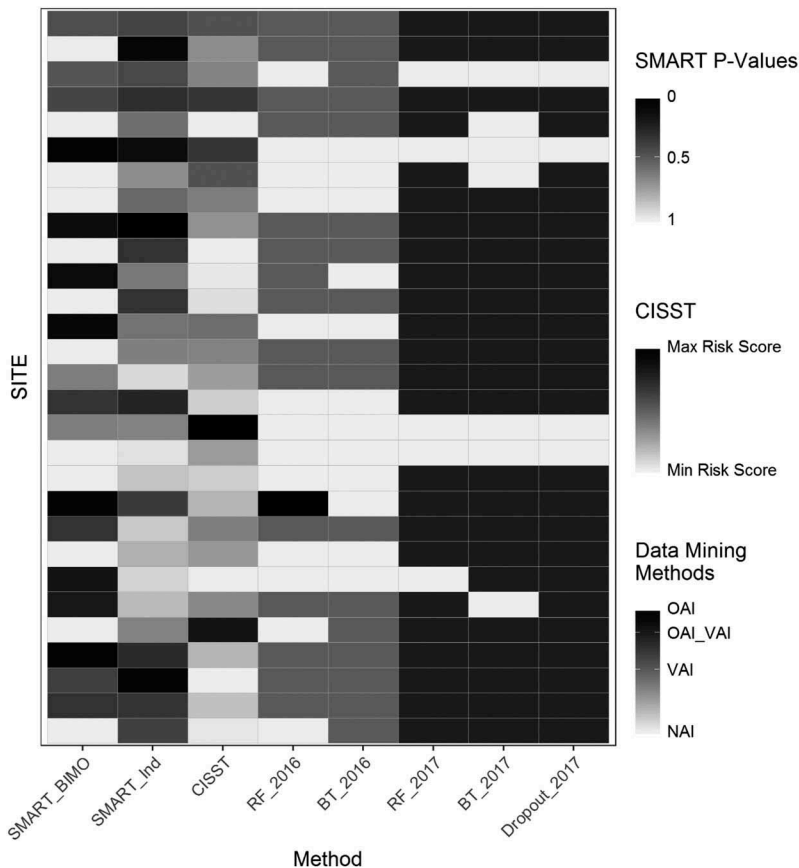


Figure 4. Heat map for study A. SMART_BIMO represents the SMART analysis using summary-level data and SMART_Ind represents the SMART analysis using patient-level data. The following abbreviations were used for the data mining methods: RF – Random Forest, BT – Boosted Tree, Dropout – boosted tree with dropout.

highest percent agreement amongst all combinations of models. The average percent agreement of concordant results aggregated by study was 64.9%. While the lowest average percent agreement (58.5%) aggregated by study occurred when comparing the random forest model from 2016 and the Boosted Tree model from 2017 (combined analysis). Performing a combined analysis or by arm analysis had a minimal effect on the percent agreement between the 2016 and 2017 models. The mean difference of the aggregate averages of percent agreement between analysis methods was 2.1%.

Examining the heat maps to determine the agreement between all methods showed the occurrence of all methods indicating OAI (or low p -value/high total risk score) or all methods predicting NAI (or high p -value/low total risk score) was rare (See Figure 4–8). Furthermore, the heat maps revealed that the by arm and the combined analysis were not always in agreement amongst the respective methods. For Study A, it appears that the Random Forest, Boosted Tree, and Boosted Tree with Dropout models from 2017 were potentially overly sensitive to predicting VAI/OAI (Figure 4). Based on the historical data of Tang et al. (2016) and Jha et al. (2017) VAI and OAI outcomes should account for approximately 40% of the predictions. The heat maps for D and E are challenging to interpret due to the large numbers of clinical sites (Figures 7 and 8).

To summarize, the agreement between the data mining models of 2016 and 2017 appears to be higher than the agreement between the SMART (Venet and Doffagne 2016) analyses p -values and the CISST risk scores. Upon examining the agreement of all tools using heat maps, we found that the

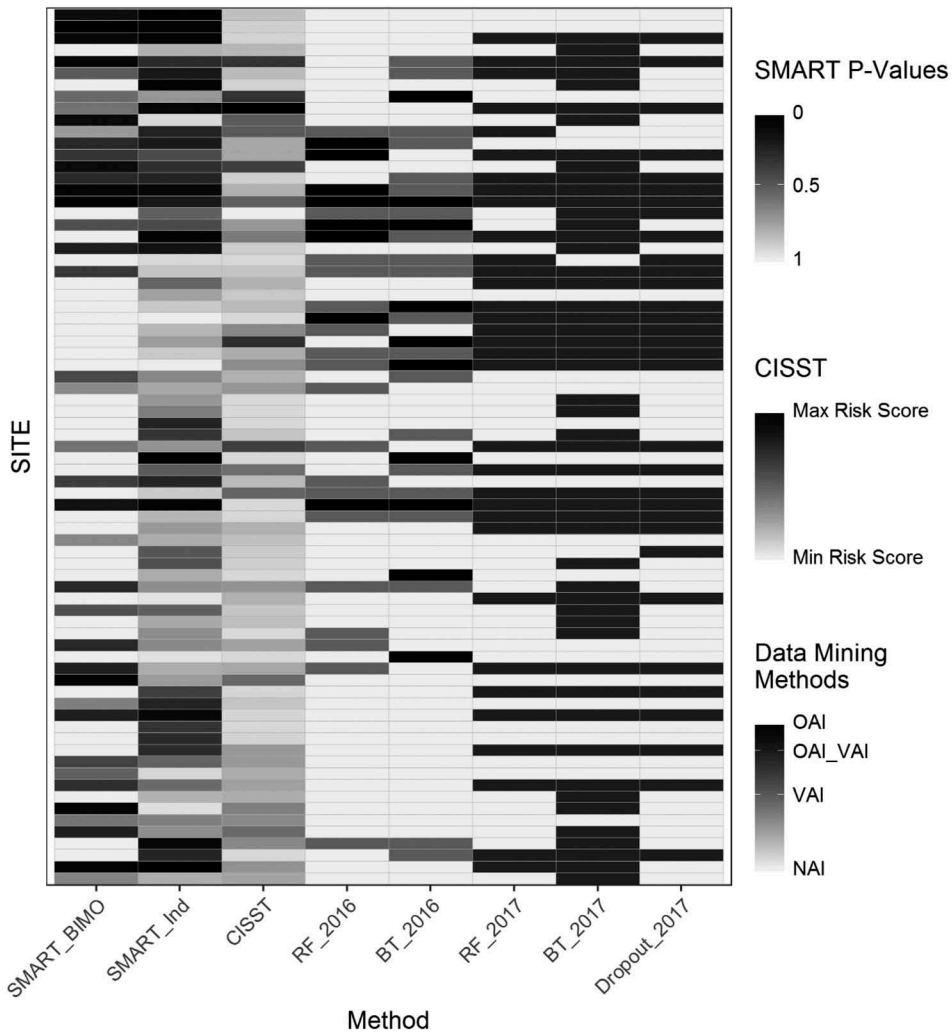


Figure 5. Heat map for study B. SMART_BIMO represents the SMART analysis using summary-level data and SMART_Ind represents the SMART analysis using patient-level data. The following abbreviations were used for the data mining methods: RF – Random Forest, BT – Boosted Tree, Dropout – boosted tree with dropout.

tools were rarely in complete agreement. The data mining methods of 2017 were most likely to identify an OAI or VAI outcome correctly; however, this seems to result in reduced specificity. The SMART (Venet and Doffagne 2016) analysis appears to have the highest specificity which results in a lower sensitivity. For the CISST and data mining models of 2016, sensitivity and specificity seem to fall somewhere in between the two extremes, i.e., high sensitivity and low specificity or high specificity and low sensitivity.

5. Discussion

This study aimed at determining the extent and degree of agreement of three different types of tools (i.e., unsupervised statistical monitoring, supervised data mining, and a risk model developed using expert opinions) in properly classifying a clinical site as NAI, VAI, or OAI. Only the data mining models of 2016 predicted NAI, VAI, or OAI. The data mining models of 2017 combined the VAI and OAI categories.

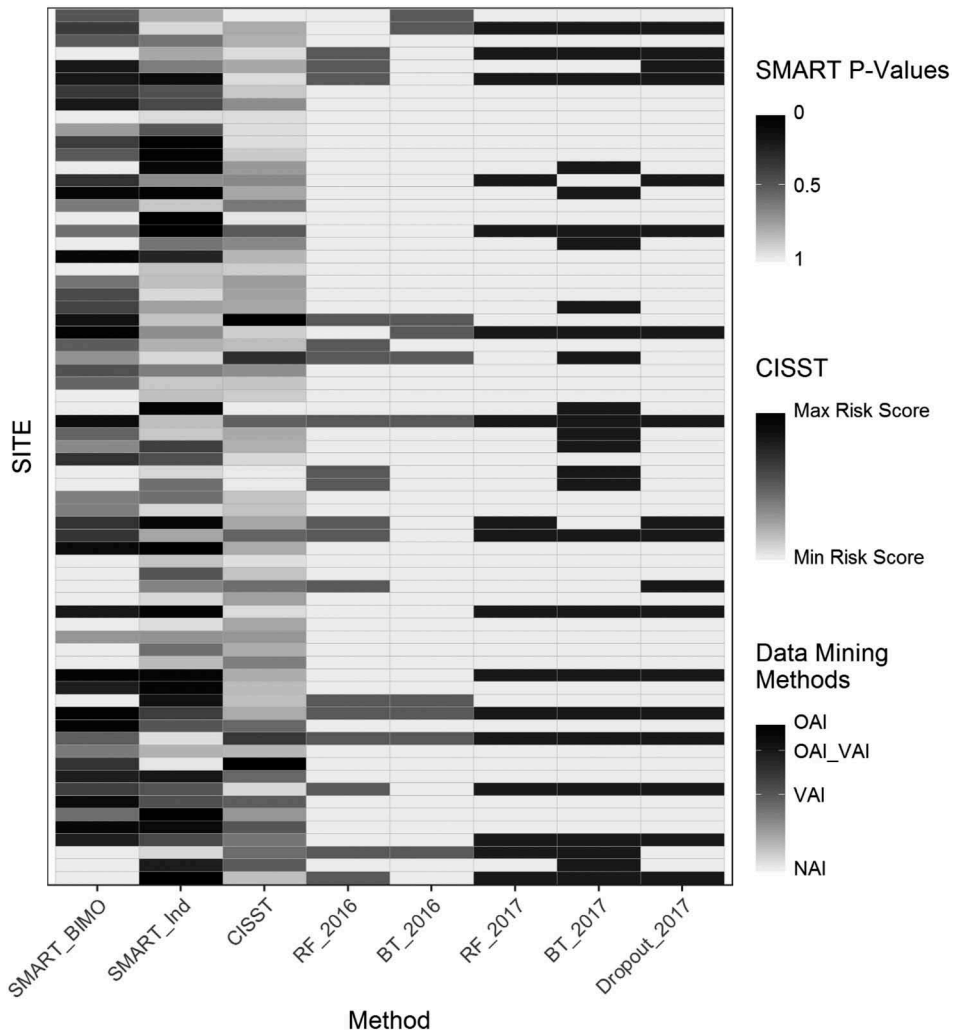


Figure 6. Heat map for study C. SMART_BIMO represents the SMART analysis using summary-level data and SMART_Ind represents the SMART analysis using patient-level data. The following abbreviations were used for the data mining methods: RF – Random Forest, BT – Boosted Tree, Dropout – boosted tree with dropout.

The SMART (Venet and Doffagne 2016) analysis outputs p -values in which an *a priori* alpha level can be used to identify atypical sites. Lastly, the CISST uses risk scores which are relative to the data being analyzed. Therefore, no global rules exist for selecting sites using the CISST. We addressed this limitation by converting all predictions to a [0,1] continuous scale and using heat maps to compare the different methods. The heat maps presented their challenges in that judgments had to be made in interpreting the predictions from the SMART (Venet and Doffagne 2016) analysis and the CISST. Furthermore, the data mining models of 2017 did not distinguish between VAI and OAI, muddying the interpretation.

Not only were the tools' outputs unique to the method, but each tool used a different set of KRIs to make predictions. Additionally, four KRIs had missing data. There were substantial amounts of missingness (near 100%) for the KRI financial disclosure. The data mining models of 2016 and 2017 require complete case analysis. These models were previously developed, so removing the covariate financial disclosure from the model was not an option. Therefore, imputation was used assuming the

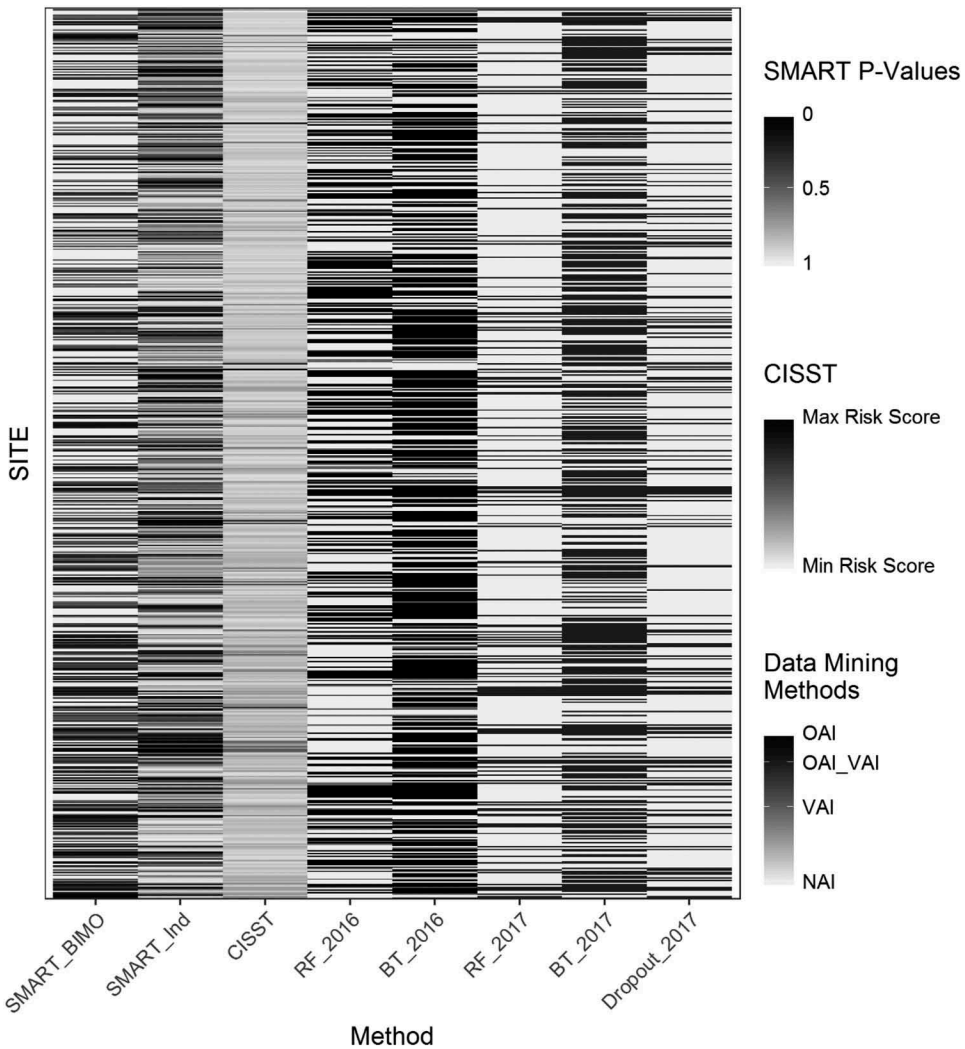


Figure 7. Heat map for study D. SMART_BIMO represents the SMART analysis using summary-level data and SMART_Ind represents the SMART analysis using patient-level data. The following abbreviations were used for the data mining methods: RF – Random Forest, BT – Boosted Tree, Dropout – boosted tree with dropout.

data were MAR. It is not possible to exclude MNAR as the mechanism of missingness. If the data were MNAR, then assuming MAR could have biased the results.

The final limitation that cannot be overcome is the possibility of selection bias. There is a non-zero probability that the data used to train the data mining models and the data used in this study had clinical sites that were included in both datasets. This could result in a spurious confirmation of the predictive ability of the data mining models.

6. Conclusions

The limited number of official outcomes precludes the uses of any hypotheses testing. However, descriptive statistics and graphical plots helped provide insights into the site selection process. Notwithstanding limitations, this study highlighted the complexities of the site selection process.

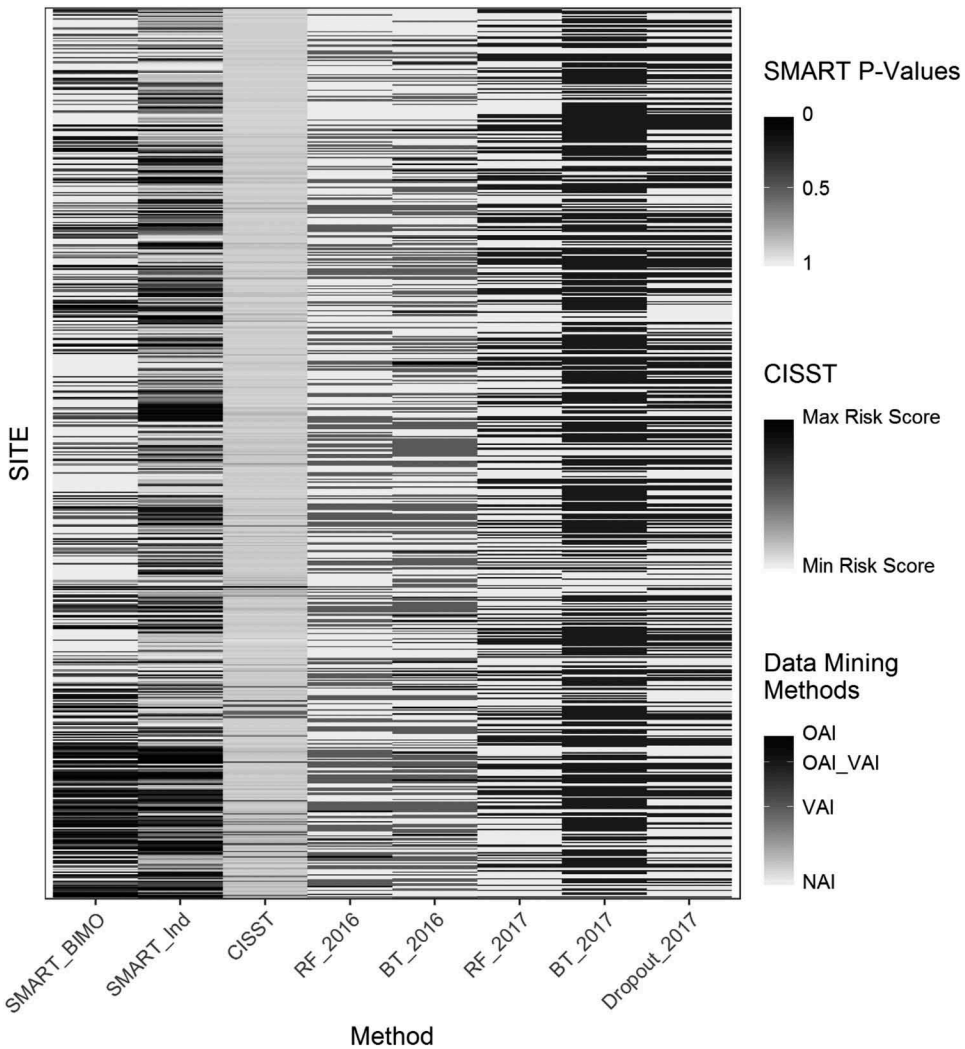


Figure 8. Heat map for study E. SMART_BIMO represents the SMART analysis using summary-level data and SMART_Ind represents the SMART analysis using patient-level data. The following abbreviations were used for the data mining methods: RF – Random Forest, BT – Boosted Tree, Dropout – boosted tree with dropout.

The agreement between methods was lower than expected. It may be that some methods perform better under certain conditions; however, these conditions are not entirely clear. Furthermore, a combination of methods may be required for a complete picture of the site selection process. Further research into the clinical investigator site selection is warranted.

Acknowledgments

The authors would like to acknowledge the assistance and insights of Michael Johnson, FDA/CDER/OTS/OCS and Jean Mulinde, FDA/CDER/OSI.

References

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16:321–357. doi:10.1613/jair.953.

- Core Team, R. 2016. *R: A language and environment for statistical computing (Version 3.3.2)*. Vienna, Austria: R Foundation for Statistical Computing.
- Core Team, R. 2018. *R: A language and environment for statistical computing (Version 3.5.0)*. Vienna, Austria: R Foundation for Statistical Computing.
- Honaker, J., G. King, and M. Blackwell. 2011. Amelia II: A program for missing data. *Journal of Statistical Software* 45 (7):1–47. doi:10.18637/jss.v045.i07.
- Jha, C., E. Rantou, and P. Schuette. 2017. Data mining tool for clinical site investigator inspection. Unpublished manuscript.
- Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. *R News* 2 (3):18–22.
- Schafer, J. L., and J. W. Graham. 2002. Missing data: our view of the state of the art. *Psychological Methods* 7 (2):147. doi:10.1037/1082-989X.7.2.147.
- Tang, M., E. Rantou, and P. Schuette. 2016. Exploring data mining methods for clinical site investigator inspection. Unpublished manuscript.
- Trotta, L., Y. Kabeya, M. Buyse, E. Doffagne, D. Venet, L. Desmet, ... K. OBA. 2019. Detection of atypical data in multicenter clinical trials using unsupervised statistical monitoring. *Clinical Trials*. Advance online publication. doi:10.1177/1740774519862564
- U.S. Food and Drug Administration. 2018a. Bioresearch monitoring technical conformance guide. <https://www.fda.gov/downloads/drugs/developmentapprovalprocess/formsubmissionrequirements/ucm332468.pdf>.
- U.S. Food and Drug Administration. 2018b. E6(R2) good clinical practice: Integrated addendum to ICH E6(R1). <https://www.fda.gov/media/93884/download>.
- U.S. FDA. Inspections database frequently asked questions. 2018c. Last Modified July 2, 2018. Accessed August 26, 2019. <https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/inspection-references/inspections-database-frequently-asked-questions>.
- Venet, D., and E. Doffagne. 2016. *Statistical monitoring applied to research trials (Version 1.9)*. Mont-Saint-Guibert, Belgium: CluePoints.
- Venet, D., E. Doffagne, T. Burzykowski, F. Beckers, Y. Tellier, E. Genevois-Marlin, ... M. Buyse. 2012. A statistical approach to central monitoring of data quality in clinical trials. *Clinical Trials* 9 (6):705–713. doi:10.1177/1740774512447898.
- Wickham, H. 2009. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. Retrieved from <http://ggplot2.org>.