Therapeutic Innovation
& Regulatory Science
2016, Vol. 50(1) 15-21

DOI: 10.1177/2168479015620248
tirs.sagepub.com

Detecting Data Quality Issues in Clinical Trials: Current Practices and Recommendations

David Knepper, MS, MBA[1], Christian Fenske, MPH[2], Patrick Nadolny, MS[3], Alun Bedding, PhD[4], Elena Gribkova, MSc[5], John Polzer, DVM, MS, MS[6], Jennifer Neumann, CCDM[7], Brett Wilson, BSP[8], Joanne Benedict, MS[9], and Andy Lawton, CSTAT[10]

Abstract

Background: Data quality issues in clinical trials can be caused by a variety of behaviors including fraud, misconduct, intentional or unintentional noncompliance, and significant carelessness. Regardless of how these behaviors are defined, they may compromise the validity of the study results. Reliable study results and quality data are needed to evaluate products for marketing approval and for decisions that are made on the use of medicine. This article focuses on detecting data quality issues, irrespective of origin or motive. Early detection of data quality issues are important so that corrective actions taken can be implemented during the conduct of the trial, recurrence can be prevented, and data quality can be preserved. *Methods:* A survey was distributed to TransCelerate member companies to assess current strategies for detecting and mitigating risks involving fraud and misconduct in clinical trials. A review of literature across many industries from 1985 to 2014 was conducted using multiple platforms. *Results:* Eighteen TransCelerate member companies anonymously responded to the survey. All of the respondents had one or more existing strategies for fraud and misconduct detection. The literature search identified current practices and methodologies across many industries. *Conclusions:* TransCelerate recommends the creation of an integrated, multifaceted approach to proactively detect data quality issues. Detection methods should include a strategy tailored to the characteristics of the study. Some sponsors are taking advantage of more advanced methods and integrated processes and systems to proactively detect and address issues, relying on advances in technology to more efficiently review data in real time. Further research is underway to assess statistical data quality detection methodology in clinical trials.

Keywords

good clinical practice, risk-based monitoring, centralized monitoring, statistical monitoring, risk indicators, clinical trial fraud, clinical trial misconduct

Introduction

Data quality issues in clinical trials are an important topic in the biopharmaceutical industry and clinical trials in particular and result from a variety of behaviors including misconduct, intentional or unintentional noncompliance, and significant carelessness. Regulators depend on the validity of study results to evaluate products for marketing approval. Reliable results are based on quality data. The US Food and Drug Administration (FDA) has described quality data as data that are fit-for-purpose and sufficiently accurate to support regulatory decisions and sponsor claims about a product and its labeling.[1] Quality in clinical trials is defined as the absence of *errors that matter*, which are errors that have a meaningful impact on patient safety or interpretation of study results.[2] The goal is not to eliminate all errors but to eliminate errors that matter. It is in this context that the unethical behaviors of fraud, misconduct,

[1] Business Operations, Allergan, Jersey City, NJ, USA
[2] Clinical Risk Management, Eli Lilly and Company, Indianapolis, IN, USA
[3] Clinical Programming and Performance Management, Bioinformatics Operations and Systems, Allergan, Irvine, CA, USA
[4] Biostatistics, Roche Products, Welwyn Garden City, Hertfordshire, UK
[5] Clinical Trial Support and Compliance, Pfizer, Moscow, Moscow, Russia
[6] Global Statistical Sciences, Eli Lilly and Company, Indianapolis, IN, USA
[7] Clinical Data Management, Roche Group, Genentech, South San Francisco, CA, USA
[8] Global Development Operations, Research and Development, Bristol-Myers Squibb, Princeton, NJ, USA
[9] Global Product Development, Roche, Genentech, San Francisco, CA, USA
[10] Biometrics and Data Management, Boehringer Ingelheim, Bracknell, UK

Submitted 14-Sep-2015; accepted 2-Nov-2015

Corresponding Author:
Andy Lawton, CSTAT, Biometrics and Data Management, Boehringer Ingelheim, Bracknell, RG40 4XF, Berkshire, UK.
Email: w.a.lawton@aol.co.uk

data falsification, and significant carelessness are important in clinical trials. These behaviors can produce errors that have a meaningful impact on patient welfare or interpretation of trial results, which may jeopardize the understanding of the risk-benefit of medicinal products marketed to the public.

In the proposed rule for Reporting Information Regarding Falsification of Data, the FDA defines falsification of data as "creating, altering, recording or omitting data in such a way that the data do not represent what actually occurred."[3] The Medicines and Healthcare Products Regulatory Agency (MHRA) of the United Kingdom takes an impact-oriented approach that focuses on misconduct "which is likely to affect to a significant degree (a) the safety or physical or mental integrity of the subjects of the trial, or (b) the scientific value of the trial."[4] The MHRA focuses more on the impact of data quality issues than the behaviors that produce them.

Misconduct and fraud (a severe type of misconduct) have generally been viewed to include a level of deception,[5,6] perpetrated to gain profit or an unfair or dishonest advantage,[5] at odds with what is agreed to be a sound scientific method, and/or generally falling short of good ethical and scientific standards.[6,7] Although fraudulent activity implies intent to deceive,[5,6] unintentional noncompliance is commonly associated with carelessness, misunderstanding of or lack of clarity in the protocol, or the inability to perform assigned tasks. In clinical trials, examples of fraud and misconduct may include activities such as fabricating patients, falsifying eligibility data to enroll otherwise unqualified patients, and knowingly fabricating data when a study procedure was inadvertently missed instead of documenting that the procedure was not done. An example of unintentional noncompliance may include an accidental error when rounding measurements.

There may be differences in the meaning and severity of the behavioral terms of fraud, misconduct, fabrication, falsification, and noncompliance; however, irrespective of how these behaviors are defined, they each produce issues with data quality and their impact can be detected through the same methods. As such, the present study approaches the detection of these behaviors and their impact in the same manner by focusing on proactively and efficiently detecting data quality issues that can jeopardize the reliability of study results. Early detection of data quality issues is important so that corrective actions can be taken, recurrence can be prevented, and data quality can be preserved.

In an effort to provide solutions for efficient and effective biopharmaceutical research and development, TransCelerate Biopharma Inc. (TransCelerate) assessed the current landscape of methods used to detect data anomalies suggestive of errors, noncompliance, or misconduct.

Materials and Methods

In the present study, a survey and a literature review were conducted to assess the current landscape for proactively identifying falsified or fabricated data.

Survey

The survey was distributed to 18 TransCelerate member companies in 2015 to benchmark their existing detection strategies (online Appendix 1). The survey responses were collected anonymously in an online survey tool and descriptive statistics were generated using Microsoft Excel.

Literature Review

A prospectively defined literature search was conducted in Medline, Embase, Biosis, Current Contents, PsycINFO, ABI/Inform, and Google Scholar using keywords and database-specific terminology (online Appendix 2). The literature that was reviewed focused on the following specific elements of data quality issues including, but not limited to, fraud, falsification, fabrication, and misconduct:

- Definitions and types
- Trends (eg, incidence and prevalence)
- Root causes and contributing factors
- Benchmarked detection methods and impact assessment

While the literature review focused on clinical research, other industries (eg, financial services and retail) were included in order to understand practices that may be relevant to the biopharmaceutical industry. The literature review included original research articles, conference abstracts, regulatory guidances, and commentary and editorials on original research published between 1985 and 2014.

Results

Survey

All 18 TransCelerate member companies responded to the survey; 16 completed the entire survey and 2 did not answer all questions. All data were included in the results. The survey explored the prevalence of detection strategies including site monitoring, auditing, data review activities, central monitoring, and advanced statistical analysis. Figure 1 illustrates the various methods used and how many companies used each method.

Site monitoring, auditing, data review activities, and central monitoring were most commonly applied across all studies, whereas advanced statistical analysis methods were used less frequently. For example, among the 14 companies using central monitoring, only 8 used it on risk-based monitoring (RBM) studies. Three of the 8 companies employing advanced statistical analysis were piloting this strategy, including the use of proprietary methods on studies using central monitoring.

Twelve of 16 companies modified their detection strategies within the past 5 years. Nine companies modified their strategies as recently as within the last 2 years. Figure 2 illustrates the main reasons that companies changed their strategies.

The survey also evaluated tactical aspects of fraud and misconduct detection (Figure 3). For the 16 responding companies, the 3 most commonly used tactics to detect fraud and

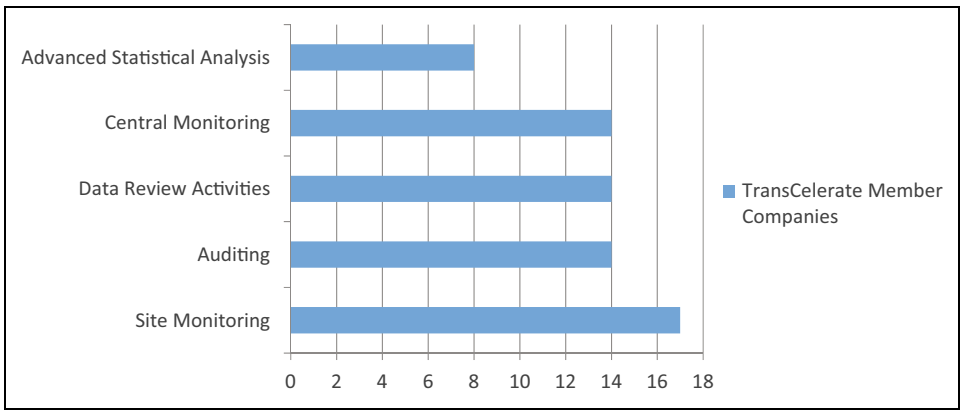

Figure 1. Strategies used by TransCelerate member companies to detect fraud and misconduct in clinical trials.

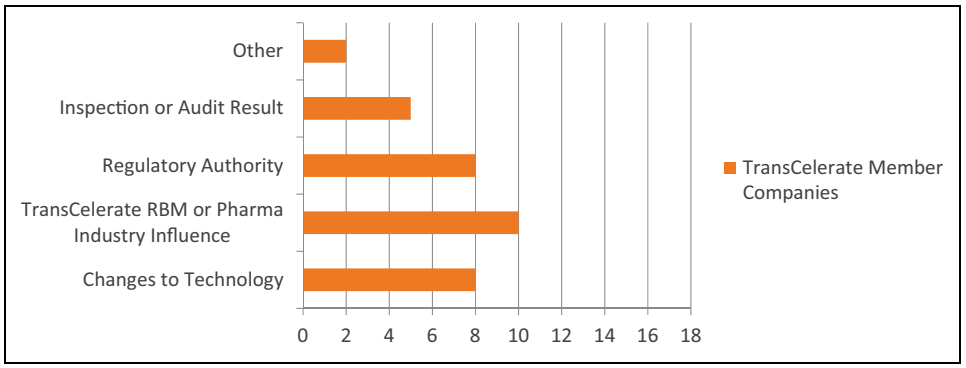

Figure 2. Key influencers of TransCelerate member companies to change strategies on detecting fraud and misconduct in clinical trials within the last 5 years.

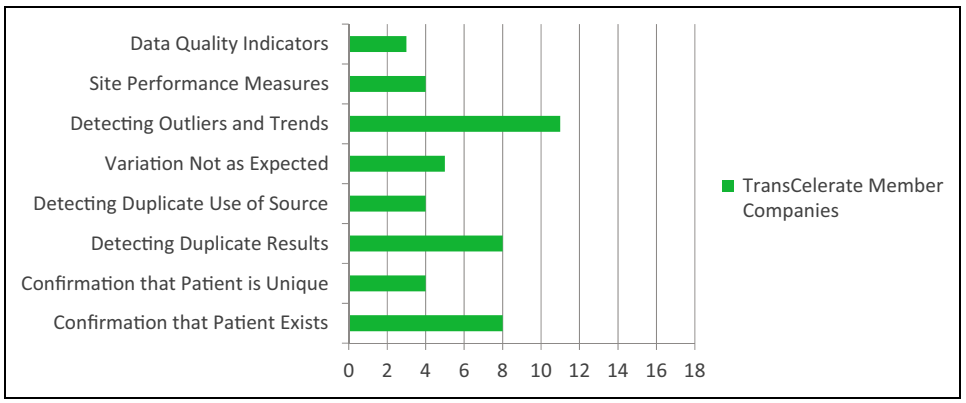

Figure 3. Most common tactics used by TransCelerate member companies to detect fraud and misconduct in clinical trials.

misconduct in clinical trials were detecting outliers and trends (11 companies), confirming that patients exist (8 companies), and detecting duplicate results (8 companies). Other tactics used less frequently included detecting unexpected variation, measuring site performance, detecting duplicate use of source data, and examining data quality indicators.

Lastly, the survey assessed the technical capabilities used in fraud and misconduct detection. The majority of respondents (14 of 16) used data analytics, 6 companies had robust data warehousing capabilities, and 1 company used cloud-based technology to execute their detection strategies.

Literature Review

The comprehensive literature review identified 151 articles that could be grouped into 3 broad categories based on topic as

related to fraud and misconduct: motivations, prevalence, and detection methods.

Motivations for Fraud and Misconduct

Reasons for fraud and misconduct vary depending on the circumstances, but some type of personal gain is generally involved. Academic researchers may be motivated by a desire for an influential publication as well as the associated prestige and financial rewards. According to Weir et al,[8] a trial with negative results does not generate name-making papers and gets little notice within the medical profession; thus, there is an incentive to falsify results. Motives of investigators participating in biopharmaceutical industry–sponsored research include increasing patient enrollment, reducing repetitive work, and reducing the reporting burden associated with clinical trials.[9] This may include study coordinators estimating pulse or respiration measurements to reduce the workload from repetitive vital sign assessments. In some cases, investigators may choose to incur a high number of screening failures to maximize revenue from screening activity.

Prevalence of Fraud and Misconduct

Many industries track incidence rates of fraud. For example, the insurance industry estimates that 20% of all insurance claims are fraudulent in some way.[10] The Crime and Fraud Prevention Bureau annual report for 2000 cited 4 main types of fraud in motor insurance and their associated levels of occurrence as "completely false claims (12%), deliberately misrepresenting the circumstances of the claim (32%), inflated loss value (39%), claiming from multiple insurers (3%), with 14% being attributable to other types of fraudulent claims."[11] In the health care industry, the National Health Care Anti-Fraud Association estimated that at least 3%, or more than $60 billion, of the US annual health care expenditure was lost due to outright fraud.[12]

According to Kirkwood et al,[13] fraud is relatively uncommon in scientific research, including nonclinical studies. In a survey of several thousand US scientists, almost 30% admitted to participating in some questionable research activity in their career, but only 0.5% admitted to "falsifying" or "cooking" research data.[13] However, self-reported data might not represent the true incidence of occurrence.

Prevalence of fraud and misconduct in the clinical research industry is not well documented; however, published literature suggests the reporting of fraud and misconduct in clinical trials is increasing.[8,14,15] Steen[16] examined articles reporting clinical trials between 2000 and 2010 and found that 1 in every 6070 articles was retracted. From 180 assessable retracted articles, there were 9 clinical trials with at least 200 participants each, and publications for 7 of these clinical trials were retracted for fraud.[13] Under the FDA Proposed Rule, the Agency estimates that it will receive 73 reports of data falsification per year across its multiple divisions.[3]

Detection Methods for Fraud and Misconduct

The insurance industry has several databases to assist in the detection of anomalous information at the claims stage.[11] The aim of these databases is threefold. First, they provide a way of verifying the information supplied by claimants. Second, they allow companies to assess whether claimants have a history of suspicious or similar claims. Third, they provide repositories for sharing information about claims histories across insurance fraud detection companies and other parties.

A framework within the banking industry takes advantage of domain knowledge, mixed features, multiple data mining methods, and a multiple-layer structure for a systematic solution.[17] The framework includes algorithms related to contrast pattern mining, neural networks, and decision forests. Outcomes are integrated with an overall score measuring the risk of an online transaction being suspicious or genuine. The approach is most effective with large volumes of extremely imbalanced data.

Data mining in the context of Homeland Security involves the use of data analysis tools to discover previously unknown valid patterns and relationships.[18] The value or significance of these patterns is then determined by a human reviewer. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output. Consequently, the limitations of data mining are primarily related to availability of data or skilled personnel rather than technology.

An objective within the realm of health care is to develop fraud detection methods and algorithms that are accurate, capable of handling the immense volume of health care data, and able to detect fraud in real time before incurring severe loss and damage.[12] Contrary to the immense volume of data in the health care industry, clinical trials are often challenged by low volume data conditions, especially during study conduct, challenging the utility of statistical methods and algorithms to detect data irregularities in real time. Central data monitoring, used in conjunction with an overall integrated monitoring plan that adapts to identified issues as a trial progresses, has the potential to increase the likelihood of detecting discrepant trends in data before trial closure and improve data quality and patient safety.[19]

In an attempt to detect data quality issues in clinical trials, certain types of data warrant closer scrutiny. Data items that frequently appear prone to error and/or falsification, fabrication, and omission include repeated measurements, patient diaries, and eligibility criteria.[9] Furthermore, data quality measures that may warrant additional examination to detect data quality issues include the following[19]:

- High or low enrollment rate
- Balanced randomization across treatment arms
- Target versus actual subject visit timing
- Variability in measurement error (expected to be similar across sites)
- Variability in measurements over time (lack of change may indicate problems)
- A large quantity of missing data

- Digit preference
- Underreporting or overreporting of adverse events

According to Li et al,[12] sophisticated antifraud systems incorporating a wide array of statistical methods are being developed for fraud detection in the health care sector. The major advantages of these systems include automatic learning of fraud patterns from data, specification of ''fraud likelihood'' for each case in order to prioritize efforts for investigating suspicious cases, and identification of new types of fraud.

In clinical trials, techniques for studying outliers, inliers, overdispersion, underdispersion, and correlations all rest upon the premise that it is difficult to invent plausible clinical data, particularly highly dimensional multivariate data. Furthermore, the multicentric nature of clinical trials offers an opportunity to check the plausibility of the data submitted by one site against data from all other sites.[9]

According to Venet et al,[20] statistical monitoring relies on the highly structured nature of clinical data because the same protocol is expected to be implemented identically at all participating sites. Furthermore, they suggest that statistical checks are powerful tools because the multivariate structure and/or time dependence of variables are sensitive to deviations and hard to mimic in fabricated or falsified data. Falsified or fabricated data, even if plausible univariately, are likely to exhibit abnormal multivariate patterns that are detectable statistically. In most cases, the main limitation to interpretation is the size of the trial. The methods to detect data errors can be applied to all trials; however, many of the methods that aim to detect fraud would be difficult to apply reliably in small trials or even larger trials when small numbers of patients are recruited within each site.[13]

Discussion

Although the prevalence of data fabrication, falsification, and other types of misconduct in biopharmaceutical industry–sponsored trials appears to be low, the potential impact may be substantial. An investigator engaging in data fabrication or falsification during a clinical trial may cause regulatory authorities to question the oversight abilities of the sponsor, jeopardizing study acceptance by regulators. The regulatory authorities could terminate a product's development or withdraw approval for a marketed drug if warranted.[21] The investigator could be subjected to substantial fines and imprisonment, if convicted, according to the US Code on Crimes and Criminal Procedure.[22] In addition, these issues may damage society's perceptions of the biopharmaceutical industry and reduce the willingness of individuals to participate in or trust the results of future clinical research.[8] However, the most immediate risk of these behaviors, as well as the primary target of detection methods, are data quality issues that may jeopardize the validity of study results.

The survey results showed that companies were taking precautions to detect such issues in data quality by increasing scrutiny in this area and increasing implementation of a more comprehensive detection approach, with some companies beginning to rely more on statistical analysis and advances in technology. All of the survey respondents reported having detection strategies in place, most companies employed these methods across all trials, and all but one company used at least 2 detection methods.

Comparing strategies versus technical capabilities demonstrated that most of the surveyed companies employing a central monitoring strategy also used data analytics and visualization technology. Specifically, 13 of the 14 companies performing central monitoring used data analytics and visualizations, with almost half also using robust data warehousing to aggregate and normalize data from various sources. Less than half of the respondents used advanced statistical analysis during the study to detect data anomalies suggestive of fraud and misconduct, indicating that the biopharmaceutical industry may be lagging behind certain industries in harnessing the power and utility of these advanced methods for real-time issue detection. It should be noted that the survey was limited to TransCelerate member companies and may not be representative of all entities sponsoring clinical trials. Other limitations of the survey included incomplete responses and inconsistency in the interpretation of the questions.

The prominent use of technology, including data mining and analytical techniques, to detect data quality issues is prevalent in certain industries, as revealed in the literature review.[11,12,17,18] The banking industry has well-developed statistical algorithms for detecting suspicious credit card activity in real time, allowing instantaneous acceptance or rejection of credit card charges. The biopharmaceutical industry is currently developing similar methods for more timely detection of data irregularities suggestive of fabrication, falsification, and other types of misconduct. One strategy discussed in the literature is central statistical monitoring (CSM) which is the use of various statistical tests during study conduct to detect data anomalies.[13,20,23] However, this strategy is under development and the published literature did not adequately differentiate between biopharmaceutical industry and academic trials and often selected data types based on motivations most commonly found in academic research. Furthermore, the researchers tested this strategy on completed databases and did not adequately evaluate the application of such methods on partial databases typical of clinical trials before database lock.[13,19] Consequently, the articles reviewed underscore the need for additional research on the application of CSM in biopharmaceutical industry–sponsored clinical trials.

Traditional methods of detecting data quality issues are useful; however, each has limitations in its ability to efficiently detect data quality issues. While on-site monitoring methods have detected numerous incidences of data quality issues, they are limited because of the nature of manual review and lack of ability to compare across sites. Random quality audits also require manual review and provide insufficient coverage in terms of the small number of sites audited. Existing statistical methods and data review activities are limited by a need for sufficient data volume, which is a challenge early in a study's life cycle. An overall strategy should leverage a combination of

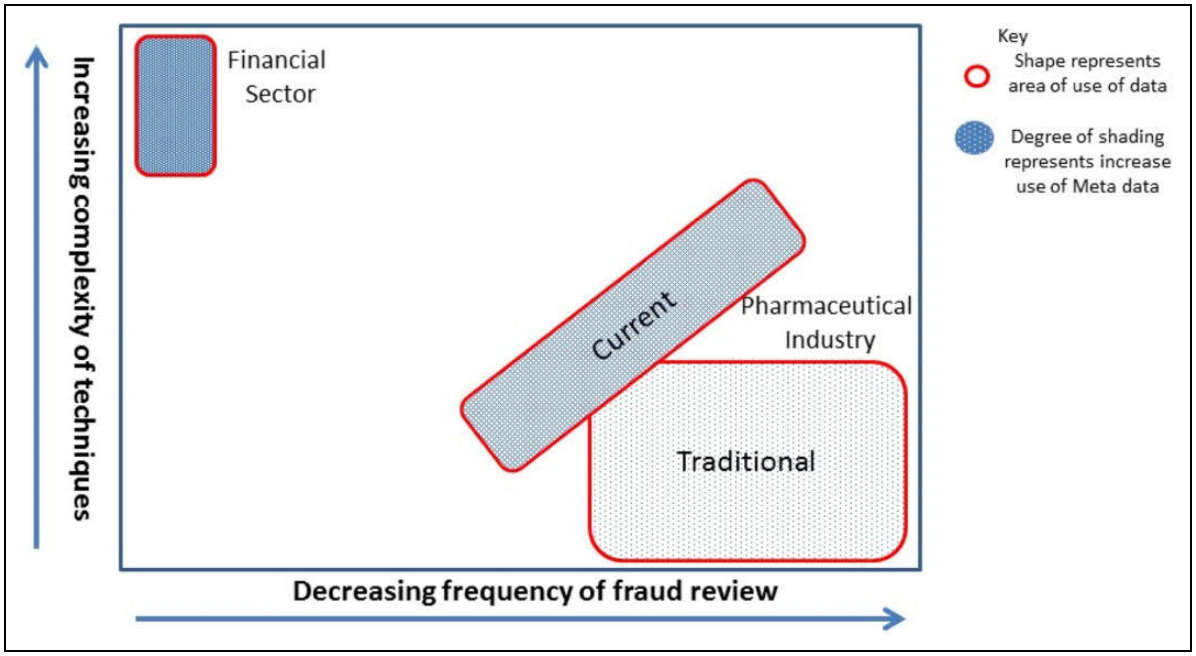

Figure 4. Illustrating industry fraud detection methods and frequency.

methods including new risk-based approaches like CSM. Advances in technology can help ease the burden on traditional manual reviews and ensure more efficient and timely identification of issues.

Conclusions

Prevention, timely detection, and mitigation of data quality issues are critical to the development of medicines and devices and the reputation of the biopharmaceutical industry. Several risk-based methods such as targeted on-site monitoring, central monitoring, and statistical monitoring can assist with detecting these threats to data quality. Although an array of detection methods can be employed, the approach should be tailored to the characteristics of the study to ensure targeted analyses and to reduce the manual burden on those overseeing data quality. For instance, programmed edit checks can help prevent accidental data errors while reducing the burden on monitoring and data review activities.

Several features of biopharmaceutical industry–sponsored clinical trials mitigate the impact of data quality issues. Randomization and blinding of treatment assignment serve to minimize bias. Occasional data errors that are random with respect to treatment, including fabrication of data, generally do "not introduce serious bias into the overall trial findings when treatment allocation is properly concealed."[24] Moreover, any data identified as suspicious are generally included in sensitivity analyses to evaluate the impact on the validity of the study results. Although the biopharmaceutical industry is currently developing more complex detection methodologies, the industry can draw on experiences from other industries, such as banking, finance, and insurance.

The discrepancy between the banking and insurance industries and the biopharmaceutical industry is pronounced (Figure 4, intended to be illustrative of how different industries currently approach data monitoring and fraud detection). This is primarily driven by the significant and direct financial losses due to fraud in the banking and insurance industries. However, there are other factors including the biopharmaceutical industry's diverse and variable sources of data requiring integration, low volume of data during study conduct, and paper-based data collection methods in some studies. As more electronic methods are utilized, advances are being made. However, the industry will continue to encounter obstacles until direct data entry with eSource or Electronic Medical Records, global data standards, and advanced data integration systems are in place. The schematic in Figure 4 represents how various industries, including the biopharmaceutical industry, currently approach detecting fraud and misconduct.

TransCelerate recommends a multifaceted, holistic approach using all available techniques to detect and mitigate clinical trial data quality issues. Detection methods should be integrated into existing practices. The analysis plan should be tailored to the characteristics of the study to ensure the most efficient and effective strategy. The goal is to eliminate errors that matter. The overarching objective of all data quality monitoring techniques is the assurance of the validity of study results.

Future Research

One of the overarching goals of TransCelerate is to provide recommendations on best practices to identify issues and to define expectations on the rates of issues affecting data quality, including fraud and misconduct. Consequently, TransCelerate

is preparing to conduct an original research project testing CSM methods on a biopharmaceutical industry–sponsored clinical trial database in conditions found during study conduct. Building on research reported in several prominent articles,[9,13,19,20,23,25,26] TransCelerate will study the utility of various statistical tests in high- and low-volume data conditions typically found during study conduct. The main focus will be the types of errors that are typically found in biopharmaceutical industry–sponsored research.[9] The study will investigate the limitations of a battery of statistical tests, including options for reporting and graphical display, and consider the best strategy for interpreting the statistical output. At the conclusion of this study, TransCelerate will publish results and discuss possible recommendations including how CSM fits into the broader quality management architecture. The cost effectiveness of different methodologies will also need to be determined to develop medicines for patients more efficiently.

Acknowledgments

The authors gratefully acknowledge the support of TransCelerate BioPharma Inc, a nonprofit organization dedicated to improving the health of people around the world by accelerating and simplifying the research and development (R&D) of innovative new therapies. The organization's mission is to collaborate across the global biopharmaceutical R&D community to identify, prioritize, design, and facilitate implementation of solutions designed to drive the efficient, effective, and high-quality delivery of new medicines. The authors would also like to thank the CRO Forum established by ACRO for their review of the draft manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Online supplemental material for this article is available on the journal's website at http://tirs.sagepub.com/supplemental.

References

1. Mulinde J.The clinical trial enterprise: defining quality. Paper presented at: DIA Quality Risk Management Conference; December 3, 2012; Philadelphia, PA.
2. Meeker-O'Connell. Update on clinical trials transformation initiative (CTTI) quality-by-design project. Paper presented at: DIA Quality Risk Management Conference; December 3, 2012; Philadelphia, PA.
3. Draft guidance on reporting information regarding falsification of data. *Fed Regist*. 2010;75(33):7412-7426.
4. United Kingdom Medicines and Healthcare Products Regulatory Agency. Guidance for notification of serious breaches of GCP or the trial protocol, Version 5, Final 060114. 2014.
5. Hamrell MR. Raising suspicions with the Food and Drug Administration: detecting misconduct. *Sci Eng Ethics*. 2010;16(4): 697-704.
6. DeMets DL. Distinctions between fraud, bias, errors, misunderstanding, and incompetence. *Control Clin Trials*. 1997;18(6): 637-650.
7. Al-Marzouki S, Roberts I, Marshall T, Evans S. The effect of scientific misconduct on the results of clinical trials: a Delphi survey. *Contemp Clin Trials*. 2005;26(3):331-337.6.
8. Weir C, Murray G. Fraud in clinical trials. *Significance*. 2011; 8(4):164-168.
9. Buyse M, George SL, Evans S, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Stat Med*. 1999;18(24):3435-3451.
10. Subelj L, Furlan S, Bajec M. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Syst Appl*. 2011;38(1):1039-1052.
11. Morley N, Ball LJ, Ormerod TC. How the detection of insurance fraud succeeds and fails. *Psychol Crime Law*. 2006;12(2):163-180.
12. Li J, Huang KY, Jin J, Shi J. A survey on statistical methods for health care fraud detection. *Health Care Manage Sci*. 2008;11(3): 275-287.
13. Kirkwood AA, Cox T, Hackshaw A. Application of methods for central statistical monitoring in clinical trials. *Clin Trials*. 2013; 10(5):783-806.
14. Steen RG. Retractions in the scientific literature: is the incidence of research fraud increasing? *J Med Ethics*. 2011;37:249-253.
15. Seife C. Research misconduct identified by the US Food and Drug Administration. *JAMA Intern Med*. 2015;175(4):567-577.
16. Steen RG. Retractions in the medical literature: who is responsible for scientific integrity? *Am Med Writ Assoc J* 2011;26:2-7.
17. Wei W, Li JJ, Cao LB, Ou YM, Chen JH. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*. 2013;16(4):449-475.
18. Seifert JW. Data mining and the search for security: challenges for connecting the dots and databases. *Gov Inf Q*. 2004;21(4):461-480.
19. Lindblad AS, Manukyan Z, Purohit-Sheth T, et al. Central site monitoring: results from a test of accuracy in identifying trials and sites failing Food and Drug Administration inspection. *Clin Trials*. 2014;11(2):205-217.
20. Venet D, Doffagne E, Burzykowski T, et al. A statistical approach to central monitoring of data quality in clinical trials. *Clin Trials*. 2012;0:1-9.
21. United States 21 CFR §312.70 b-e.
22. Crimes and Criminal Procedure. 18 USC §1001.
23. Pogue JM, Devereaux PJ, Thorlund K, Yusuf S. Central statistical monitoring: detecting fraud in clinical trials. *Clin Trials*. 2013; 10(2):225-235.
24. Baigent C, Harrell FE, Buyse M, Emberson JR, Altman DG. Ensuring trial validity by data quality assurance and diversification of monitoring methods. *Clin Trials*. 2008;5:49-55.
25. O'Kelly M. Using statistical techniques to detect fraud: a test case. *Pharm Stat*. 2004;3(4):237-246.
26. Wu X, Carlsson M. Detecting data fabrication in clinical trials from cluster analysis perspective. *Pharm Stat*. 2011;10(3):257-264.