

# Advanced Analytics for the Monitoring of Clinical Trials

Moving Beyond Risk-Based Monitoring



Jennifer Dennis-Wall, PhD  
Mujataba Sharief, PhD  
Hemalatha Raju, PhD



## Introduction: Innovative Approaches to Data Sciences

Datasets are one of the most valuable commodities of the 21st century. Retailers, manufacturers, and even government agencies are increasingly turning to data-centric approaches that rely on large datasets to make better and more informed decisions, run leaner operations, and drive bottom-line revenue.

Clinical research generates a massive amount of data. Every step in the clinical trial process—including the planning, execution, analysis, and closing phases of studies—creates enough information to completely fill multiple hard drives. Unfortunately, because of the way most clinical data programs currently work, much of these data are not used to their fullest extent; they are simply filed away.

But, what if there was a way to harness this data and turn it into a strategic advantage? Plenty of other traditional industries have found ways to leverage their customer and operations data into more efficient business processes, and clinical trial operations can do the same.

For starters, clinical trial data scientists can:

1. Create more efficient, real-time data collection employing centralized monitoring and CDISC standard datasets<sup>1</sup>
2. Provide superior analytics beyond those used for standard risk-based monitoring (RBM)
3. Implement ongoing, sophisticated, and accurate monitoring reports to quickly identify issues and correct errors

This white paper will focus on the benefits provided by a unique and advanced analytical approach to study monitoring beyond the minimum necessary RBM approach.

## Risk-Based Monitoring is Not Enough

Historically, investigators and sponsors have been responsible for ensuring that two primary factors are monitored and maintained throughout the clinical trial process: participant safety and study integrity. To protect participant safety, adverse reactions to study interventions need to be immediately reported to investigators, sponsors, review boards, and even sometimes regulatory authorities.

To protect the study's integrity, errors, negligence, and even fraud by investigators all need to be quickly identified so that corrective actions can be taken. If these two main objectives are not met when monitoring, a study can be shut down, or data from the study can be denied upon submission to the appropriate regulatory authority (e.g. Food and Drug Administration [FDA]).

Source data verification (SDV) was originally, and still is, used as a method to ensure that all data are transcribed accurately from source to case report forms (**Table 1**). For SDV, study monitors must physically crosscheck the data entries for accuracy. While this is important for study data that are transcribed from paper into multiple systems, this method will only catch transcription errors and is unable to identify other types of issues such as inter-site or inter-rater variability and measurement errors. Additionally, this type of monitoring is labor-intensive, costly, and prone to human error – a problem especially when studies are large, multicentered, or international.

Coverage & Visit Requirements	Traditional Monitoring	Risk-Based Monitoring	"Big Data" Monitoring
On-site source verification	✓	✓	✓
Centralized source verification	✓	✓	✓
Most serious risks	✓	✓	✓
Most likely risks	✓	✓	✓
All other data	✓	✗	✓
Fraud detection	✗	✗	✓
Inter/Intra-site analytics	✗	✗	✓
Inter/Intra-subject analytics	✗	✗	✓
On-site requirement	100%	Reduced	Reduced

**Table 1.** Capabilities of different clinical trial monitoring approaches throughout history.

In an effort to help sponsors and investigators understand what to expect and how to design their studies in order to meet requirements, the FDA issues guidance on study monitoring. The most recent guidance<sup>2</sup>, published in August of 2013, outlined in detail an RBM approach designed to minimize time spent monitoring while identifying the most important and likely errors. The RBM approach recommends that a risk assessment<sup>3,4</sup> is conducted prior to the start of a study to predict these key issues and risks. Risks can (and often do) include compliance rates for procedures that are difficult to follow or the accuracy and precision of subjective study endpoints. These anticipated risks are then prioritized for monitoring, with other risks being deemed less important.

In addition to this more targeted RBM approach, the guidance suggested a partial transition from on-site monitoring to centralized monitoring where possible in order to reduce the number of costly and time-intensive monitoring visits. The guidance also suggested SDV of only a sample of study documents rather than checking through all source documents for accuracy.

Overall, the objective of the RBM guidance was to reduce the amount of work for sponsors monitoring a study. However, these new guidelines also present new challenges for sponsors and their CROs because the RBM approach can overlook important variables or outcomes that are not part of the predetermined risk assessment and monitoring plan. Additionally, in order to monitor from a central location, the burden now rests upon the site to transfer large amounts of source documents to a central repository so that a remote team can cross-verify the source data with those entered into the electronic data capture (EDC) systems. Generally, this process also involves a fair amount of redaction to protect sensitive subject data, yet another labor-intensive task which is only now being facilitated by automation technology available on the market. Finally, in order to conduct SDV on only a sample of data, sponsors must be confident that the sample is representative of the rest of the dataset. In either case, if any data discrepancies are missed during the SDV process, they may not be identified until the end of study when there is no opportunity to take corrective action.

The FDA makes it very clear that the guidance is simply a recommendation, allowing for a lot of interpretation on their part when it comes time to approve or deny a New Drug Application (NDA). If a study has unexpected issues that are deemed to affect the integrity of the study, those issues may still interfere with the FDA's approval of a drug or device even if they weren't initially identified as posing a large enough risk to be monitored with the RBM plan.

The bottom line is that RBM is too targeted and provides a limited scope of the study's quality and integrity. Even though RBM addresses the cost concerns of traditional monitoring, it does so at the expense of the complete oversight of study data. There should not have to be such a tradeoff between cost and quality; what is now needed is an approach that maintains or even potentially increases the oversight and quality of data, while at the same time reducing the cost of monitoring.

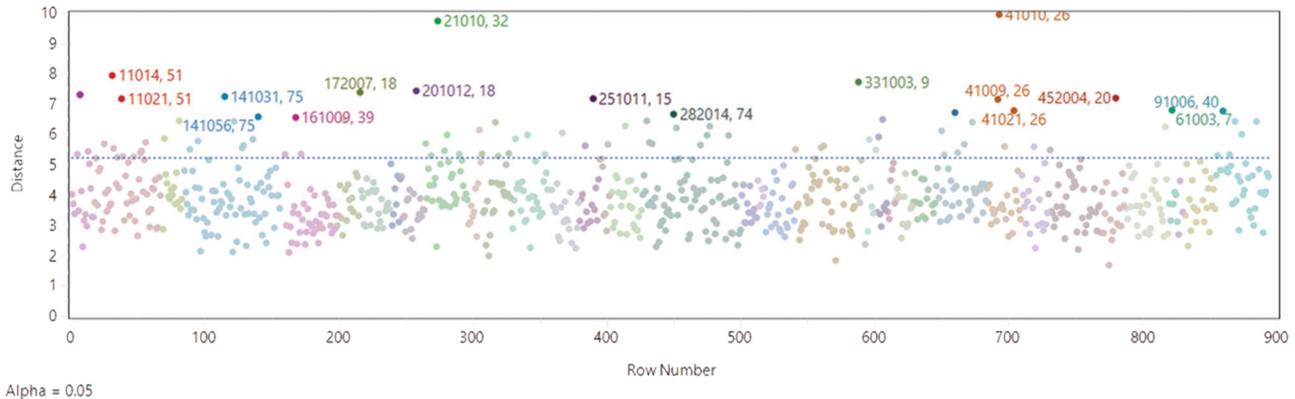
## Lessons Learned from Big Data – Efficient Data Processing

In the search for a more holistic, all-encompassing use of study data, we turn to the strategies of big data analysis. The term “big data” was coined in the 1990's and refers to datasets that have **volume** (datasets have significantly more observations than before), **velocity** (data stream in from sources at higher speeds in order to keep up with volume), and **variety** (data come from many different sources and in different formats). These aspects of big data have posed new challenges for data managers and statisticians. Large volumes of data necessitate better infrastructure: bigger servers, faster data transfer capabilities, and more processing power. On top of that, statistical software must be able to compute the larger volumes of data within reasonable time frames. To accommodate these modern needs, better technology and analytical methods have been developed to enable parsing and analyzing of big data.

Cloud-based storage has revolutionized how data is managed and stored. Now that datasets don't have to take up as much physical space, the limit to how much data can be collected has increased. Computing power has also improved to accommodate large volumes of data; processing power has increased by a trillion-fold over the past 60 years<sup>5</sup>.

Statistical packages now use more efficient processes and advanced algorithms to be able to compute larger datasets. When planned for appropriately, the many variables recorded during data collection for a clinical trial can be analyzed automatically because of CDISC standard file formats and validated statistical packages designed to use those standards. This in turn saves data scientists a lot of time and sponsors a lot of money. Visualization of the data has also improved to be able to see what these larger, more complex datasets look like (**Figure 1**) and to determine how they should be analyzed.

Although these approaches were developed out of necessity for big datasets, they can apply to the analysis of smaller datasets as well to make any type of analysis more efficient. Because of these advanced analytics, all of the data collected during a study can quickly and easily be used for the purposes of monitoring a study's quality and integrity. In this way, unexpected anomalies in the datasets are discovered more quickly and brought to the attention of study sponsors and monitors.



**Figure 1.** A scatterplot depicting Mahalanobis distances for multivariate data. Individual data points contain information on multiple variables; this type of analysis and accompanying visualization identifies outliers (numbered, above the dotted cut-off line) in a dataset by comparing the distributions of many variables at once. The variables included in this type of analysis can be any variable collected during a study and can therefore identify anomalies in any dataset.

## A Smarter, Big Data Approach to Monitoring - Expanding Beyond RBM Analyses

As stated previously, the two main objectives of monitoring a clinical trial lie in preserving study data integrity and protecting the patient. To protect patient safety, adverse events and abnormal lab values need to be closely monitored in real-time. Study data can be compromised in multiple ways.

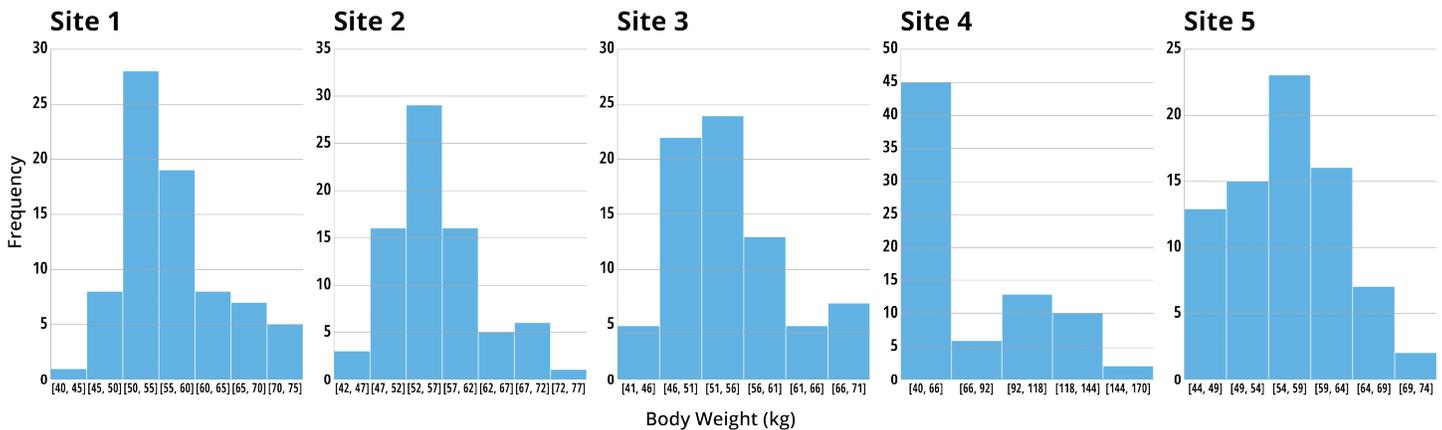
The problem with the RBM approach is that it is an oversimplified solution to a problem that is not simple. Errors can originate in any of the data – not just the data that is deemed “important” or “critical” in the study’s preliminary risk assessment. So, the approach of just simply not looking at a portion of the data in order to reduce time requirements is inherently risky because issues that manifest in that “unimportant” or “noncritical” data are completely missed. In other words, even noncritical data can tell you a lot about the quality of the data in general and the quality of the important data specifically. Monitoring approaches that analyze all of the data in a “big data” way are being adapted to fulfill these needs of monitoring clinical trials comprehensively yet efficiently.

## Comprehensive Coverage

Any time an investigator makes an honest, or even a negligent, mistake when entering study data into source documents or the EDC, it can turn into a big problem. Less common but more severe is when investigators or even study participants knowingly commit fraudulent actions; this can happen when investigators fabricate study data in order to receive compensation for enrolling study participants or when patients enroll at multiple sites for the same study. Fortunately, approaches now exist that can identify any errors from mistakes to intentional fraudulent actions by making use of all of the study’s data.

One of the simplest and yet most eloquent ways to quickly identify errors in data entry is based on the principle that distributions of study variables should naturally be very similar between different study sites. A variable that exhibits a normal distribution at all but one study site would raise suspicion (**Figure 2**). For example, this approach has been used to identify instances in which values of weights were accidentally entered in pounds instead of kilograms, a likely occurrence in an international study that includes American sites not accustomed to using metric units.

Patients are often compensated well for participating in clinical trials. For this reason, some are tempted to travel between study sites posing as multiple different participants, and a site investigator would be none the wiser. Duplicate dates of birth can be identified with certain analytical tools, and because patients cannot drastically change their biometrics, it is rather easy to discern whether two people with the same date of birth are actually different people.



**Figure 2.** Results from an analytical simulation to detect errors in data entry. Sites 1, 2, 3, and 5 all tend to exhibit a normal distribution for body weight (kg) while site 4 exhibits a skewed distribution. This scenario could result from errors in recording weights using the wrong units. In this situation, a query would be raised to the clinical monitors for further investigation.

In extreme cases, investigators may commit fraud in order to boost their enrollment numbers and subsequently receive more compensation from the sponsor. One clever way to detect fraud employs the concept of digit preference. Digit preference is the humanistic phenomenon of preferring certain numbers, such as 1's and 0's, to others. When people fabricate data, they tend to enter numbers they prefer. The prevalence of digits can be compared between sites to identify digit preference (**Box 1**).

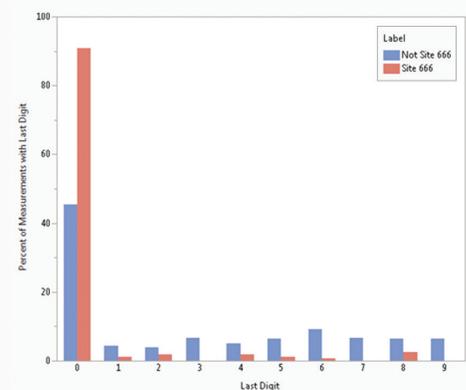
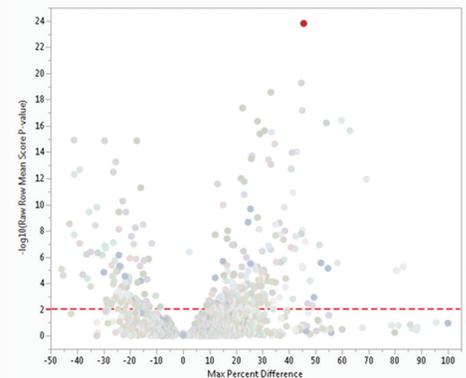
These data discrepancies are easy to detect using a comprehensive, big data-style analysis of all variables. It would not be so easy, if even possible, to do this as a monitor simply verifying source data. What's more, with on-site monitoring, the time lapse between the error's occurrence and the review of the data could be weeks, months, or even years.

**Box 1 | A case study of investigator fraud employing digit preference.**

There was something fishy about the data coming from Site 666, but the data scientists couldn't quite place it. The investigator was entering mass amounts of past-due records into the EDC at once and enrolling participants in batches rather than gradually as would be expected. A query was raised and the data sciences team investigated further.

Because fraud was suspected, it made sense to investigate whether data was being fabricated, and digit preference was suspected as the underlying detectable behavior. When the last digit of every biometric value was analyzed by site by creating a volcano plot (panel A, depicting the magnitude of the change by significance), an outlier site, Site 666, was identified. The last digits of all values from Site 666 were then modeled in a histogram (panel B), and it was clear that 0's and 1's were entered more frequently for biometric values than other values. This would not be expected with real-life biometric measurements.

In response to the suspicious activity, a thorough investigation of Site 666 was immediately launched and monitoring visits to the site were increased. Visits revealed that the values entered into the EDC by the investigator did not match the source data. Because the reports were generated in real-time while the study was happening, there was still time to add additional sites to replace site 666 if it was deemed necessary.



## Efficiency

A standardized approach to data collection, such as one using CDISC standards, focuses on efficiency and ease of use by allowing for the automation of many analytical processes, including those used for monitoring. Automation of analytical processes means that reports can be generated nearly instantly, allowing for data-informed decisions to be made at any point during a trial. Adverse events and abnormal lab values can thus be reported immediately and monitored carefully using the centralized monitoring system. Sponsors are thus able to more closely adhere to GCP and ethical standards because of the immediate identification of adverse events and deviant data. Other errors in the data are detected quickly, and corrective actions can take place much earlier in the study than when monitoring relies on in-person visits to detect errors.

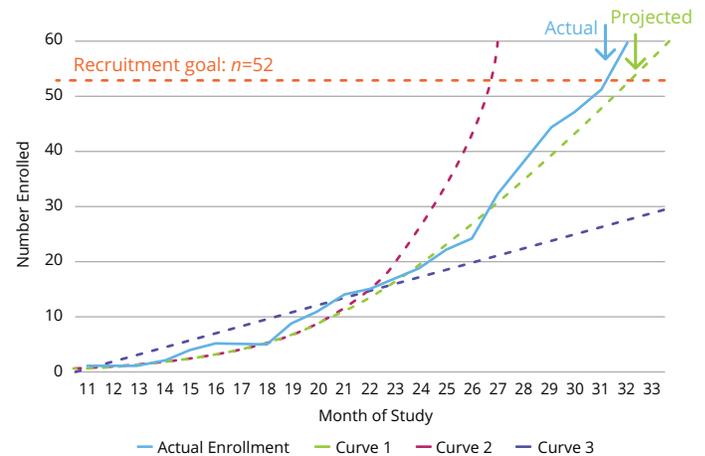
Not all studies are identical, though. For that reason, automated processes need to be adapted to each study, requiring the efforts of a talented data sciences team. When done correctly, monitoring, and even other phases of a study such as recruitment, can benefit from these efficient analytical processes.

## Additional Uses for Big Data Analytics - Recruitment

Recruitment is a notoriously challenging phase of many trials, and a lot of pressure is put on the study team to meet recruitment goals. When goals are not met and recruitment must be extended, milestones are pushed and costs are increased for the sponsor. A few extra months of recruitment can add millions of dollars to the budget. Even worse, a study can stall completely if there are not enough funds available to extend recruitment.

A more comprehensive approach to data collection means that sponsors and CROs can leverage data collected from previous studies and use it to predict how long recruitment will take, allowing for more accurate budget and timeline planning from the start. After recruitment starts and a few months of data points have been collected, a predictive curve can be generated (**Figure 3**) to see whether recruitment is on target. If not, adjustments to recruitment budgets and outreach strategies can be made earlier in the study to get recruitment timelines back on track. Study data can also be used to identify optimal sites for a study based on past performance and other quality metrics. In these cases, the more data, the better.

Comprehensive use of study data makes for more accurate budgets and timelines, monitoring above and beyond FDA requirements, and identifying errors and fraud quickly. All of this protects the integrity of the study and helps to mitigate risk.



**Figure 3.** Projected versus actual enrollment for a recent trial. The most realistic predictive curve (Curve 1) was generated at month 23 to determine whether additional sites needed to be added to fulfill recruitment goals ( $n = 52$ ). Data from month 11 through month 23 were used to generate the curve. Neither a simple exponential curve (Curve 2) nor a linear curve (Curve 3) would have predicted enrollment time requirements as accurately as the method employed by Biorasi's data sciences team.

## Conclusion

The big data approach towards clinical trial data monitoring is not yet an industry standard, but it is the future of more accurate and efficient data analysis and monitoring. Biorasi employs a more unique and thorough approach to analytical techniques in order to streamline recruitment and efficiently monitor studies beyond what is capable with RBM. To improve our monitoring capabilities even further, Biorasi is working on projects that include the design of algorithms to generate weekly monitoring reports with a single click. Our approach gives us the best of both worlds: the ability to ensure quality across all the data collected in a clinical trial while at the same time minimizing the low value activities that drive up study budgets.

Future white papers will detail the types of monitoring reports that can be created using the analytical approaches discussed here and expand upon the way big data can be used to drive recruitment performance. Sponsors will be able to see how adding more comprehensive data analytics processes to their trials can identify small opportunities for improvement that translate into much larger cost savings and efficiencies.

## References

1. Biorasi White Paper. Putting CDISC Standards to Work. Available at <https://www.biorasi.com/project/wp-putting-cdisc-standards-work/>. Accessed 6/22/2018.
2. Guidance for Industry: Oversight of Clinical Investigations - A Risk-Based Approach to Monitoring. U.S. Department of Health and Human Services, Food and Drug Administration. August 2013. Available at <https://www.fda.gov/downloads/Drugs/Guidances/UCM269919.pdf>. Accessed 6/25/2018.
3. IEC 31010:2009 Risk Management – Risk Assessment Techniques. International Organization for Standardization (ISO). Available at <https://www.iso.org/standard/51073.html>. Accessed 6/26/2018.
4. Risk-based monitoring in clinical trials. JMP. Available at [https://www.jmp.com/en\\_us/software/clinical-data-analysis-software/risk-based-monitoring.html](https://www.jmp.com/en_us/software/clinical-data-analysis-software/risk-based-monitoring.html). Accessed 6/22/2018.
5. Processing power compared. Experts Exchange. Available at <https://pages.experts-exchange.com/processing-power-compared>. Accessed 6/25/2018.

## About the Authors



### Jennifer Dennis-Wall, PhD

#### Lead Scientific Writer, Marketing

Dr. Jennifer Dennis-Wall received her PhD in nutritional sciences from the University of Florida and joined the Biorasi team shortly after. Jennifer cultivated interests in scientific writing and biostatistics while conducting multiple randomized trials researching various health outcomes. She now contributes regularly to the Biorasi blog.



### Mujataba Sharief, PhD

#### Associate, Data Sciences & Biometrics

Dr. Mujataba Sharief received his PhD in biochemistry from Florida International University and studied cell signaling mechanisms. He now uses his expertise to alert Biorasi's study monitors of anomalies in clinical trial datasets and to efficiently prepare data for analysis. He is currently developing Biorasi's big data analytics strategy to support trial recruitment.



### Hemalatha Raju, PhD

#### Associate, Data Sciences & Biometrics

Dr. Hemalatha Raju joined the Biorasi team after earning her PhD in human genetics and genomics from the University of Miami Miller School of Medicine. In her doctoral program, she developed and implemented computational algorithms to detect dark matter transcription across human tissues. Hema is now developing new, more efficient algorithms for Biorasi's on-study analytics and advanced risk monitoring platform.



**CLINICAL RESEARCH OPTIMIZED™**

*Powered by TALOS™*

Biorasi is a contract research organization (CRO) widely recognized for delivering success in complex clinical trials. This is possible through TALOS™, an innovative operating model that unifies systems and teams with a powerful project management methodology to ensure high quality delivery. Overall, Biorasi balances time, cost and risk to optimize the delivery of clinical studies.

Global biopharmaceutical companies have come to depend on Biorasi to deliver their most complex studies. The company's expertise includes a range of molecule types, development phases, therapeutic areas, geographies, and development programs. Biorasi has collaborated with sponsors to enable FDA, EMA, and multi-venue approvals for numerous small molecules and biologics. Biorasi, headquartered in Miami, Florida, maintains office-based teams around the globe. The company has received the coveted CRO Leadership Award from Life Science Leader magazine and has placed on the Inc. 500 list of America's fastest growing companies.

For more information, visit [www.biorasi.com](http://www.biorasi.com)



**Biorasi Global Headquarters**

19495 Biscayne Blvd. Suite 900 · Miami, Florida 33180

786.388.0700 · [info@biorasi.com](mailto:info@biorasi.com)