Using Investigative Analytics to Speed New Drugs to Market

Investigative analytics can ensure effective on-site monitoring, data integrity and compliance – while accelerating clinical trials.

Executive Summary

The clinical trial, in which a drug or device is tested for its safety and efficacy, all too often becomes an obstacle on the road to market that healthcare companies and patients cannot afford.

A shortage of new blockbuster drugs and the expiration of patents on older drugs together leave pharmaceuticals companies struggling to maintain sales and profits. Meanwhile, regulatory agencies, particularly in the U.S. and Europe, are increasingly relying on new analytic and information tools to meet pressure to approve breakthrough drugs more quickly. The more quickly and efficiently pharmaceuticals companies can conduct clinical trials, the sooner patients can receive life-changing treatment, and the sooner manufacturers see a return on their R&D investments.

Today’s time from molecule to market averages 12 years and can be painful to endure (see Figure 1). But before any such analysis can be brought to bear, a common but troublesome problem must be solved: Assuring the correctness, completeness and integrity of clinical data. Data quality is vital in any analytics exercise, but especially in a clinical trial, where the lives and health of patients require the highest possible accuracy of critical information such as a patient’s age, blood pressure, dosages administered and outcomes. This white paper describes a statistical approach to data quality monitoring and explains how it could speed clinical trials and thus bring new lifesaving treatments to market more quickly and inexpensively.

Needed: A Faster, Less Expensive Route to Data Quality

Sponsors of clinical trials have long struggled to find and correct data discrepancies that result from, among other reasons, miscalibrated devices, human error and deliberately manipulated records. This involves applying a variety of metrics and rules on data from multiple repositories. In the process, dependencies among the data sources grow exponentially; and more often than not, sponsors are left with open questions about the effect of such dependencies on data quality. These questions might include:

- Does a discrepancy in a particular variable signify a manual error or an intentional fabrication of data?
- How does a discrepancy in one variable at one site affect the overall trial results?
- What are the audit requirements to check such data?
Such questions also make it difficult to capture actionable information about data quality in a rigid set of KPIs and canned reports.

Errors in data may be introduced due to incorrect trial design or interpretation of results, procedural errors, faulty equipment, negligence or fraud. Monitoring typically involves comparing information recorded in the case report form (CRF) with the corresponding source documents through on-site visits. Such a comparison finds discrepancies resulting from transcription errors from the source documentation to the CRF, but may miss errors present in the source documents. It is also expensive and may not find data problems caused by negligence or fraud. In fact, drug manufacturers spend up to one-third of their clinical trial budgets in such labor-intensive activities.

A better approach would be to perform “adaptive” monitoring where the sites that require costly on-site visits are chosen based on key performance parameters, such as percentage of fabricated data, incorrect records, missed compliance or other serious events at the level of patient, visit or site. The use of automated analytics- and logic-based workflows, alerts, escalations and audit trails identify data quality issues and provide consistent and traceable “actionable outcomes” that reduce risk and costs while improving quality and compliance.

Analytics In-Depth

Analytics is the open-ended search for patterns, anomalies and clusters - i.e., clues - that can be used to formulate questions or which can be correlated with events, conditions or phenomena. Investigative analytics allows users to ask a series of quickly changing, iterative questions to understand why something did or did not happen and how to optimize a particular outcome in the future, resulting in deeper and richer insight. It can also be used to describe the output of a test.

There are two types of analysis most applicable to improving data quality - exploratory data analysis and inferential statistics.

Exploratory Data Analysis

Exploratory data analysis (EDA) emphasizes the substantive understanding of data, creating graphic representations of data, using robust measures and subset analysis and taking a skeptical, flexible approach on which methods to use in assuring data quality. One frequent product of EDA is analogies that help identify suspicious outliers or extreme values, or that present the data distribution in a scatter plot as, for example, an ellipse, horseshoe or straight line. All this...
EDA can identify the areas of greatest concern, pinpointing areas for further analysis and moving the decision-maker closer to a decision.

Examples of how EDA can be used at the patient and site level include:

- **Outlier analysis** identifies those observations that deviate from the majority of the data values, thus signaling possible data quality issues. These may be hard to detect and may be innocuous if their frequency is low. Such data values may be common to only a certain section of the trial (e.g., laboratory data) and are randomly distributed. Box plots, histograms and scatter diagrams are very helpful in visualizing these types of data values.

  Depending upon the data distribution, the analysis can be carried out using various techniques such as the 2SD and 3SD methods, Tukey’s Method (1.5IQR and 3IQR)\(^2\), adjusted box plot and median rule. These techniques can be used to statistically/mathematically confirm the graphical findings of the EDA.

  When seeking additional structure in univariate distributions or when a number of distributions need to be compared, a box plot is often used. The box plot offers a five-point summary in schematic form (see Figure 2).

  The box plot compares all clinical trial sites, identifying those that show abnormal values for a particular variable. This helps identify sites having discrepancies due to manual error, fabrication of data or individual bias.

  - **Repeated value analysis** is especially useful for uncovering data that has been fabricated or manipulated to magnify the effectiveness of a drug. It does so by examining the variability in the data, using graphic representation to check for suspicious patterns or frequencies of particular values. Values that are repeated more often than expected can be further checked for randomness through a run test, a statistical procedure that determines whether a sequence of data is truly random.

  For example, after standardizing the values of different lab tests that use different units of measurement, various methods such as a histogram or scatter plot (see Figure 3) can showcase variations in frequencies of different values for patients at each site. The intermittent peaks at specific sites may point toward data fabrication or other discrepancies at a site. Run tests can provide strong evidence for data having been manipulated or fabricated, and even for patients having been invented to strengthen the trial results.

**Understanding Outlier Analysis: Box Plot for Sites**

![Figure 2](image-url)
• **Principal component analysis** (PCA) is a data reduction technique that transforms a large dataset into manageable form by plotting data with more than three variables into two dimensions. Figure 4 illustrates how this approach can be applied to find sites that show strong evidence of irregular behavior through two dimensional diagrams of two principal components. (Principal components are the linear combination of different exact levels of significance obtained through a t-test by comparing averages of different variables of a particular site versus those of all other sites in a particular clinical trial.)

**Inferential Statistics**

Inferential statistics is an open-ended activity that looks for patterns, anomalies and clusters that can be used to formulate questions or correlate with events, conditions or phenomena. It answers questions such as: “What will happen

**Principal Component Analysis in Action**

![Figure 4](image-url)
after we cleanse the data using a particular EDA technique?” and “How should we check the inter-dependencies in dependent and independent variables?”

This analysis is often a natural extension of EDA, driven by curiosity about the future and whether observed trends or patterns will continue. The answers provide insight into impending outcomes, which lets the user take corrective action before any harm results from poor-quality data, providing a “best” or “preferred” course of action.

For example, if data fabrication issues are identified at a particular site, the organization can refer the site for an immediate audit. Confirmation on the potentially problematic nature of those centers could then be obtained and appropriate steps taken to rectify the problems and avoid any penalties in terms of cost and opportunity. Examples of inferential statistics include regression analysis, time series analysis, structural equation modeling and association rule techniques.

Examples of the use of inferential statistics in improving drug trial clinical data include:

- **Using the condition index methodology to perform audits on only the most troublesome sites.** Sites that justify an audit can be identified by measuring each on parameters such as its percentage of fraudulent or missing data. This can be conducted through composite indexing of various dimensions on a multisite trial. These dimensions include:
  - **Average value:** A t-test can compare the average values of a dimension under study for a particular site compared to others.
  - **Variability:** An F-test can be an effective way to test homogeneity in sites (i.e., one site’s variability compared to that of all other sites).
  - **Fraudulent or fabricated data:** A Z-test can help identify which sites have an unusual level of suspicious markers such as missing values, outliers or adverse/severe events.
  - **Frequency distribution:** A Chi Square test can showcase a distribution of repeated values for all sites. Since there could be cases where someone will enter the same dummy data for a particular variable each day, a high frequency of repeated values at a site is an indicator of possible fraudulent data.

These tests can compare information on one dimension (the column vector of the exact level of significance) of a particular site against all other sites. The findings of all the tests for all the dimensions can be combined and the summary used as a composite index, which ideally can handle multidimensional issues. Together, these tests can make the auditing process more targeted, better informed and more efficient.

Figure 5 depicts a sample composite index, where the number of variables considered, the number of Z-tests applied and are all the sites where trials have been conducted. Using the results of all tests on different variables of each site, we can create an index that can be used to score/rank these sites. We have used red, yellow and green to identify the sites with high, medium and low risks of problematic data.

Of course, identifying all relevant variables that make a site problematic or nonproblematic is difficult. However, if all the relevant variables needed for a model are available, simple logistic or multinomial logistic regression can be applied

### Composite Index Scoring: An Illustrative Approach

<table>
<thead>
<tr>
<th>Sites</th>
<th>Character under study 1</th>
<th>Character under study p</th>
<th>Index</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t-Test</td>
<td>F-Test</td>
<td>Z-Test 1</td>
<td>...</td>
</tr>
<tr>
<td>S1</td>
<td>0.00 0.00 0.03</td>
<td>0.76 ...</td>
<td>0.00 0.00 0.02</td>
<td>...</td>
</tr>
<tr>
<td>S2</td>
<td>0.01 0.01 0.08</td>
<td>0.94</td>
<td>1.00 0.01 0.07</td>
<td>...</td>
</tr>
<tr>
<td>S3</td>
<td>0.00 0.01 0.12</td>
<td>0.63</td>
<td>0.00 0.01 0.42</td>
<td>...</td>
</tr>
<tr>
<td>Sn-1</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00 0.00</td>
<td></td>
</tr>
<tr>
<td>Sn</td>
<td>0.09 0.04 0.89</td>
<td>0.03</td>
<td>0.00 0.04 0.30</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5
to predict problematic sites. This is one way to check for the variables that are the biggest contributors of possible discrepancies on a site.

- **Using conditional probabilities to predict patient health.** In any clinical trial, patient health status is of the utmost importance, and is often checked at each visit with the help of a quality of life (QoL) questionnaire, which measures a person's sense of well-being stemming from satisfaction or dissatisfaction with the areas of life that are important to them. Applying the answers to QoL questionnaires, rating/responses provided by different patients at different visits can be analyzed using conditional probabilities to predict their responses during the next visit. The predicted responses can be compared to the actual responses, and used to help patients take precautions to improve their health.

If enough information on factors impacting patient health exists, other techniques such as logistic regression (simple/ordinal/multinomial) can also be used to predict health status. Figure 6 highlights the patient's current visit health status and next visit expected health status. P1, P2, ..., Pm are m patients and V1, V2, ..., Vv denotes v visits of a patient during the trial. The individual scores for a particular question, for a visit and for each individual patient are recorded. A combination of these scores is then used to predict the expected health score for future visits.

**Looking Forward: Next Steps**

Investigative analytics can improve both on-site monitoring and data quality, providing a more economical route to compliant, cost-effective clinical trials. Leveraging automation and increasing access to information, workflows and alerts drives improvements in quality and compliance.

As described in this white paper, effectively using various analytics capabilities at different stages of the trial process allows investigators to address a wider set of decisions in greater detail and create a culture of data-driven decision-making. Therefore, sponsors should use investigative analytics in combination with other techniques, such as clustering, decision tree and support vector machine techniques, to bring the business process into more complete control and to address a wider array of business problems that are part and parcel of any compliant, accurate and intelligent clinical trial design.

Investigative analytics using advanced statistical methods can detect deviations in data patterns. Translating data discrepancies into quality checkpoints optimizes site visits and assures data quality, thereby reducing the cost, effort, risks and time involved in clinical trials.

**Conditional/Probable Patient Health Status**

<table>
<thead>
<tr>
<th>Patients</th>
<th>Current Health Status</th>
<th>Expected Health Status at Next Visit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V1</td>
<td>V2</td>
</tr>
<tr>
<td>P1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>P3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pn-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pm</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6**
References

• Ferrans, C., and Powers, M., “Quality of Life Index: Development and psychometric properties,”
• Oscar Podesta, Rafael Diaz, Peter Sandercock, et al., “Sensible approaches for reducing clinical trial
• Joel H. Pitt and Helene Z. Hill, “Statistical Detection of Potentially Fabricated Numerical Data: A
  Case Study.”
• Al-Marzouki Sanaa, Evans Stephen, Tom Marshall and Roberts Ian, “Are these data real? Statistical
  methods for the detection of data fabrication in clinical trials.”
• Lisa Kart, Alexander Linden, W. Roy Schulte, “Extend Your Portfolio of Analytics Capabilities,”
  Gartner report, September 23, 2013.
• Jonathan R Emberson, Douglas G Altman, et al., “Ensuring trial validity by data quality assurance
  Clinical Trials Transformation Initiative.”
• Rachelle A. Fong, Rita C. Purvis, et al., “Risk-based Monitoring Strategies for Improved Clinical Trial
  Performance.”
• Härdle W, Simar L., Applied Multivariate Statistical Analysis.
• “Risk-Adapted Approaches to the Management of Clinical Trials of Investigational Medicinal
• Annett R.D., Bender B.G., Lapidus J., et al., “Predicting children’s quality of life in an asthma clinical
  trial: what do children’s reports tell us?”
• “Guidance for Industry Oversight of Clinical Investigations: A Risk Based Approach to Monitoring.”
  (Draft Guidance), U.S. Food & Drug Administration, August 2011.

About the Authors

Dinesh Kumar Pateria is a Manager within Cognizant Analytics Practice. Focused on life sciences, Dinesh
has eight years of experience in the analytics space with demonstrated expertise across a multiplicity
of statistical techniques and statistical models (linear and nonlinear). Dinesh holds a Ph.D. degree in
statistics from the Indian Agricultural Research Institute (IARI). He can be reached at DineshKumar.
Pateria@cognizant.com.

Sanjay Bagga is a Senior Associate within Cognizant Analytics Practice. He has six years of experience
working with various domains of healthcare and life sciences performing sales forecasts, activity opti-
mization and ROI analysis. Sanjay provides guidance, training and key interpretations in operational
analytics. He can be reached at Sanjay.Bagga@cognizant.com.
About Cognizant Analytics

Within Cognizant, as part of the social-mobile-analytics-cloud (SMAC) stack of businesses under our emerging business accelerator (EBA), the Cognizant Analytics unit is a distinguished, broad-based market leader in analytics. It differentiates itself by focusing on topical, actionable, analytics-based solutions coupled with our consulting approach, IP-based nonlinear platforms, solution accelerators and a deeply entrenched customer-centric engagement model. The unit is dedicated to bringing insights and foresights to a multitude of industry verticals/domains/functions across the entire business spectrum. We are a consulting-led analytics organization that combines deep domain knowledge, rich analytical expertise and cutting-edge technology to bring innovation to our multifunctional and multinational clients; deliver virtualized, advanced integrated analytics across the value chain; and create value through innovative and agile business delivery models. http://www.cognizant.com/enterpriseanalytics.

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world’s leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 50 delivery centers worldwide and approximately 171,400 employees as of December 31, 2013, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.