BioVisualization: *Enhancing Clinical Data Mining*

Even as many clinicians struggle to give up their pen and paper charts and spreadsheets, some innovators are already shifting health care information technology into a new paradigm. Researchers, the greater pharmaceutical industry and other stakeholders are analyzing huge amounts of aggregated information — big data — to elucidate patterns that remained hidden under old data models. Blending biostatistics, bioinformatics, computer programming and operational research, big data is expected to transform the process of clinical decision making. Of course, much of this data will come from clinical trials.

Accurate and timely data management begins with detailed and proven processes. By combining these processes with state-of-the-art data management platforms, researchers can ensure the delivery of clean data in accord with exact specifications. The ever-increasing volume of clinical and laboratory data represents a substantial resource that can provide a foundation for the improved understanding of disease presentation, response to therapy and health care delivery processes. Data mining supports these goals by discovering, unraveling and, in some cases, anticipating similarities and relationships between data elements in large datasets. Currently, medical data poses several characteristics that make the application of these techniques difficult, although there have been notable medical data mining successes. Future developments in integrated medical data repositories, standardized data representation and guidelines for the appropriate research use of clinical data will decrease the barriers to mining projects.

DATA VISUALIZATION

Large, complex datasets are generated throughout the course of a clinical trial. Biovisualization platforms are tools that enable effective data mining and ease of data interpretation. Understanding the underlying trends within data is vital to making critical decisions and accelerating time to market. However, data analysis can be challenging, time-consuming and tedious.

Increasingly, in-house biostatisticians — especially within large pharmaceutical companies — are being asked to undertake complicated and time-consuming exploratory analyses or to look at data in "different ways." Researchers want to know how to partition the demographics; they want to think differently about the data so they can stratify their populations and, perhaps, formulate a hypothesis for what might be a potential biomarker for a new drug. Another key driver behind developing these biovisualization platforms was the need to understand clinical data in real time (during the trial), which allows sponsors and moderators to make informed decisions more quickly and efficiently. During the course of a clinical trial, vast quantities of numerical data are generated. Traditionally, that information was deposited into a huge collection of vast spreadsheets and manually assessed and analyzed. Quite literally, researchers and scientists were left to stare at long columns of numbers and try to make sense of the trends within. Now, technology provides a better way.

SPOTTING THE OUTLIER

The rapidly increasing amount of data in biological research, experiments and clinical trials calls for effective data analysis techniques. Visual analytics techniques have proven to be an effective way to analyze biological data, enabling researchers and trial sponsors to combine the strength of automatic methods with the expert knowledge of the analyst. So, the benefit of having a biovisualization tool is to be able to go beyond the numbers and look at the visual representations that can be created with the data, which allows reviewers to find, for example, any individuals within a subject group who are demonstrating out-of-specification (OOS) results or outlier measurements that might require further investigation or intervention. They would also be able to identify any unanticipated trends within a certain population.

With visual data analytics techniques, researchers can take advantage of the raw feed data, explore it further and get a deeper understanding of that information to motivate novel hypotheses — and also to monitor the data to make sure that nothing abnormal is happening.

THE TOOLS AVAILABLE

Many providers have recently introduced visual data analytics platforms and found that a one-size-fits-all approach to biovisualization has its limitations. For example, bioinformatics experts may find a pre-programmed platform too restrictive, but, for a biologist or scientist at the bench, it can be too complicated. Consequently, when additional bioinformatic-type analysis needs to be performed, scientists often bypass visual data analytics platforms and go directly to their in-house resource.

Customizability is an important aspect of the tool: The data that the researchers are compiling is unique, and they should have access to a platform that's flexible enough to exploit it. Although data mining has been around for some time, the scale of the datasets has become enormous. Intuitive, highly interactive computational tools that perform analyses and visualize unknown trends, patterns and outliers offer a way to identify buried opportunities and risks and provide a competitive advantage. These tools accelerate the decision-making process and facilitate the identification of non-compliant results.

FUNCTIONALITY

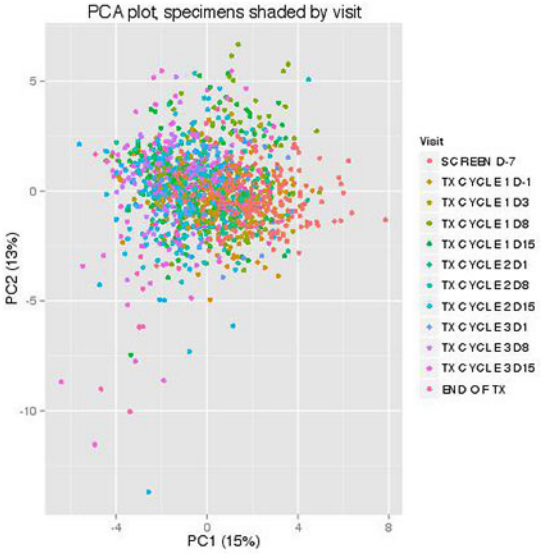
Biovisualization tools can help researchers go beyond a simple pie or bar chart to delve further into the data. Increasingly, researchers are using these platforms for two key reasons: to spot potential problems before they become major issues (from data quality or medical monitoring perspectives, for example); or to answer scientific research questions regarding biomarker evaluation or the integration of study-specific information, such as dose cohort or toxicity metrics. Information from individual subjects can be displayed, for example, to review the analyte data from a specific person in the context of the whole population to provide answers to these questions.

Furthermore, demand has been rising for more modular, customizable biovisualization tools, such as those that can be adapted so research personnel can access a very specific type of figure in a more meaningful format. Depending on the necessary type of analysis required, there are different options to choose from.

PRINCIPAL COMPONENT ANALYSIS (PCA):

The goal of this technique is to utilize all the analyte measurements to find, if possible, hidden structure in the data by minimizing redundancy and maximizing signal strength. Variables (analyte measurements) are transformed into new variables called principal components (PCs): The first principal component (PC1) accounts for the highest percentage of variance in the data; the second principal component (PC2) represents the second highest percentage of variance and so on. PCA plots (PC1 versus PC2) allow users to determine study factors (e.g., visit, investigator site, dose cohort, age group) linked to variability in analyte measurements and can often reveal unseen patterns. Figure 1 is an example of a PCA plot with the data points shaded by visit. At a glance, variability in study measurements linked to visit is expected by observing the clustering of the points along PC1. The far right section of the plot is populated by mostly orange and yellow points, which correspond to visits at the beginning of the study. The center section is populated mostly by green and light blue points, which correspond to visits in the middle of the study. The far left section is populated mostly by purple and dark blue points, which correspond to visits at the end of the study.

Figure 1: A PCA plot with the data points shaded by visit.



It should be noted that PCs do not represent single analyte measurements. Instead, they are linear combinations of them weighted by the analytes' respective loading scores derived from the PCA. Thus, loading scores (Figure 2) can be checked to better understand each analyte's contribution to PCs, keeping in mind the following points:

- The greater the loading score magnitude (positive or negative) for a particular PC, the more that analyte's measurements influence that PC.
- The closer to zero the loading score is, the less that analyte's measurements influence that PC.
- Analytes that co-localize on the loading plot are expected to exhibit similar behavior.

Inspecting Figure 2 reveals that hemoglobin (HGB), red blood cells (RBC), and hematocrit (HCT) have large PC1 loading scores (outlined in yellow), indicating these analytes strongly influence PC1. Since data variability linked to visit is expected from the clustering pattern along PC1 (Figure 1), these analytes' measurements are anticipated to vary markedly by visit. Also, these analytes are expected to exhibit similar measuremet patterns, as they co-localize on the loading plot (Figure 2). Taken together, PCA (Figure 1) and loading (Figure 2) plots empower the user to rapidly identify study factors and analytes of potential clinical significance that may deem more detailed investigation.

One way to examine analyte measurements in detail is with boxplots, either for the whole population (Figure 3) or separated by demographic (e.g., investigator site, dose cohort, age group). Each box represents the middle 50% of the data, with the horizontal line within it representing the median. Hence, the boxplots in Figure 3 provide both trending and distribution information by visit. The red blood cell (RBC) measurements (Figure 3) exhibit a marked downward trend over time, with the observed visit-linked variability consistent with the observations in the PCA (Figure 1) and loading (Figure 2) plots. The horizontal red lines represent the upper and lower reference limits, providing context to the results.

An example of a PCA-based workflow involves clinical personnel first consulting PCA plots to determine factors associated with variability in the data, then inspecting loading plots to identify analytes of interest, and last checking relevant boxplots to learn about analyte trends and distributions. Research scientists may conduct this workflow centered around dose cohort or disease subtype. Such exploratory analysis may reveal biomarkers underpinning differences between these groups, or confirm expected results. Medical monitors could rapidly identify demographics and analytes with potentially alarming trends. Project managers may be interested assessing investigator site performance. This workflow could help identify sites associated with distinct data patterns that deem follow up investigation (e.g., are differentiating factors linked to sample mishandling?).

Figure 2: Hemoglobin (HGB), red blood cell (RBC) and hematocrit (HCT) measurements have large loading scores and co-localize on the loading plot (outlined in yellow), strongly influencing PCs and showing similar behavior.

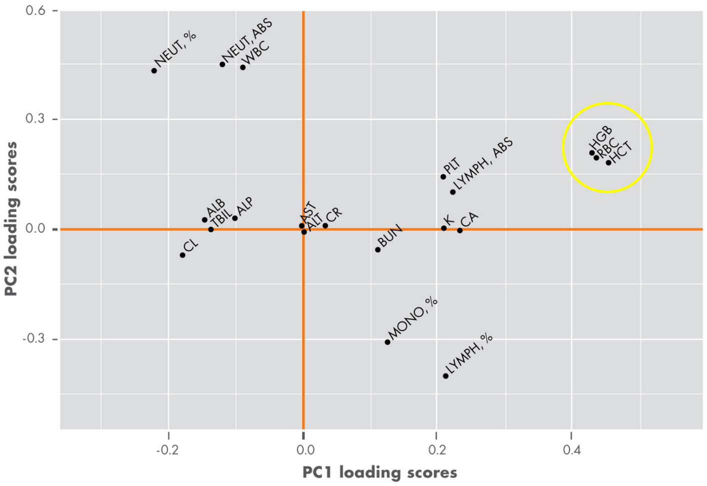


Figure 3: A boxplot displaying the red blood cell (RBC) measurements of a study population with respect to visit, allowing the user to observe trends with ease.

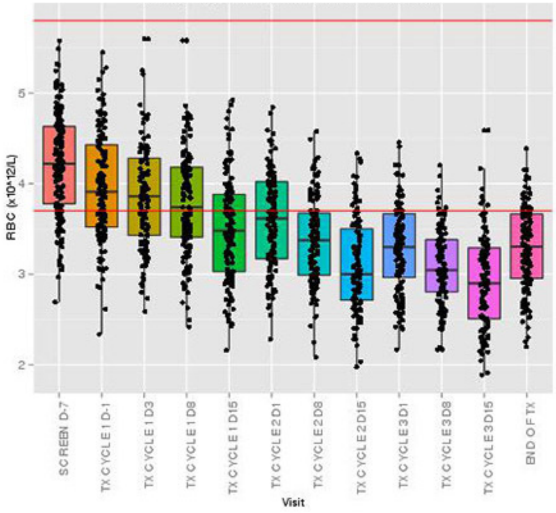
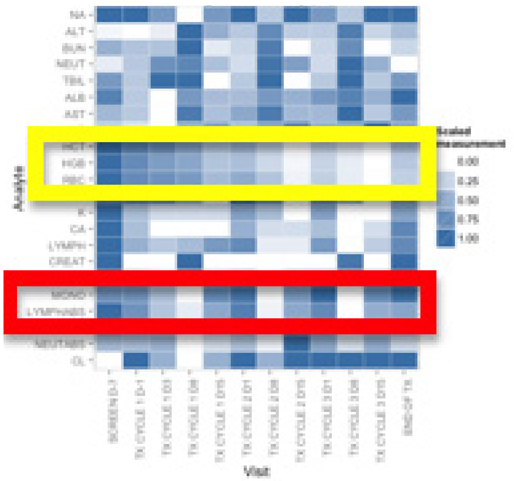


Figure 4: Heatmap of visit-specific median analyte measurements after scaling (whole study population).



HEATMAP:

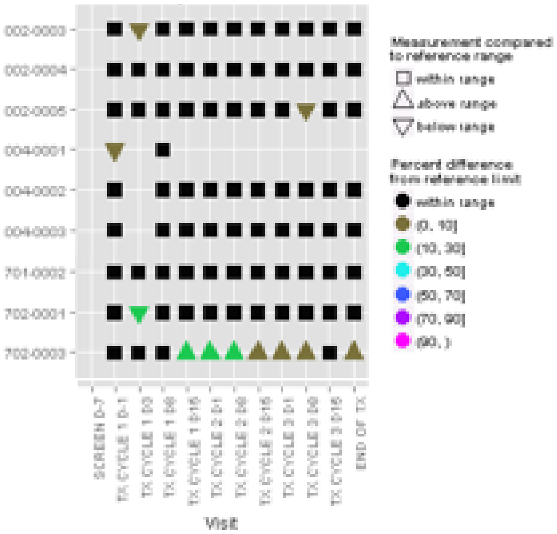
In the case of Figure 4, the purpose of a heatmap is to identify studywide patterns in analyte measurements over time, whereby the color of each cell is proportionate to the study population's median analyte measurement (0 = white, 1 = deep blue) for a particular visit after scaling. For example, the hematocrit (HCT), hemoglobin (HGB) and red blood cell (RBC) measurements (Figure 4, outlined in yellow) are relatively high at the beginning of the study and drop off with time. Other measurements, such as monocytes (MONO) and absolute lymphocytes (LYMPHABS) (Figure 4, outlined in red), tend to be oscillatory with time, because they start out at an intermediate level and then spike downward, return to the intermediate level and then spike downward again, etc. Heatmaps are useful to research scientists and medical monitors for rapidly confirming expected trends (e.g., analyte responses to drug treatment) or discovering unexpected trends. Heatmaps could also complement the workflow described in the previous section.

REFERENCE RANGE COMPARISONS:

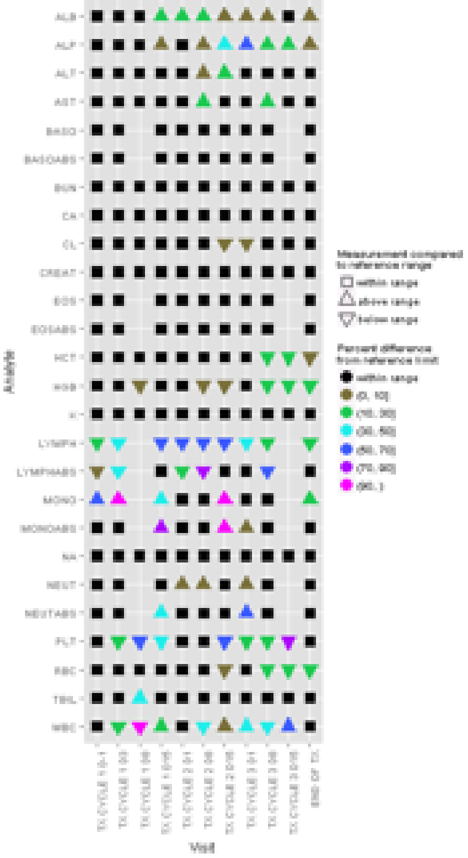
An area of concern in clinical trials is how analyte readings compare with acceptable reference ranges. These figures allows users to assess comparisons in three ways: first as a summary for a specific subset of the whole population (Figure 5a); then, as a subject profile (Figure 5b) and, finally, in a single-analyte-single-subject specific way (Figure 5c). Figure 5a is an example of a summary for a specific analyte (albumin (ALB)) and demographic (males). Rows and columns represent subject IDs and visits, respectively. Each entry conveys information based on the colors and shapes described in the legend. Black squares mean that corresponding measurements are within the analyte's reference limits, while upward and downward facing triangles indicate measurements above and below the upper and lower reference limits, respectively. The colors of the triangles represent the percent differences between the measurements and the breached reference limits. Figure 5b is an example of a summary for a specific subject. Rows represent analytes and columns visits, with entries conveying the same information as Figure 5a. Figure 5c is a plot of measurements for a single analyte from a specific subject. The horizontal black lines represent the upper and lower reference limits, providing context to the results.

These figures enable medical monitors to rapidly assess demographics, subjects, or analytes with potentially alarming trends. For example, in Figure 5a, subject 702-0003 jumps out because the corresponding measurements are consistently above the established reference interval. This subject can be further explored by examining a subject-specific profile (Figure 5b), noting that the lymphocyte (LYMPH) and platelet (PLT) results may seem interesting because they are consistently below their respective reference intervals. In either case, further information can be extracted from the data by examining the individual's records in more detailed plots like Figure 5c.

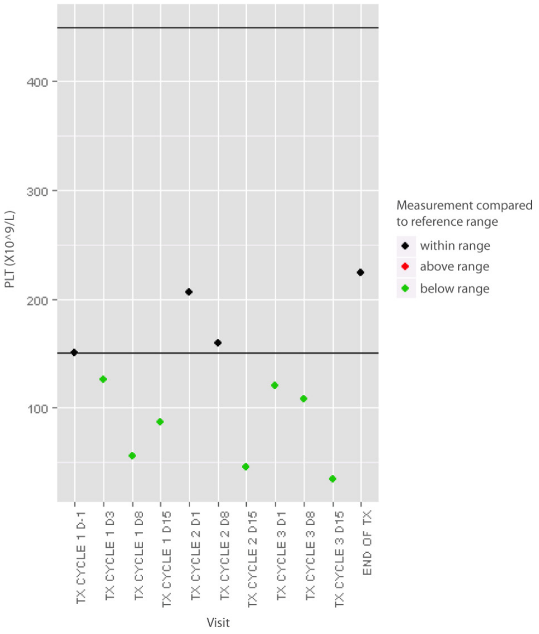
Figure 5a: A reference range comparison of male subjects.



Figures 5b: Reference range comparisons for an individual subject.



Figures 5c: A single analyte from a single subject.



BENEFITS

Biovisualization platforms and services are particularly useful for large and, perhaps more immediately, smaller pharmaceutical companies that might not have the staff, time or resources to undertake the extent of data analysis required to safeguard the ongoing process of their clinical trial. Irrespective of the labor and costs, small, medium and large product developers — that may not have access to teams of in-house biostatisticians — unfortunately often resort to spending more time than necessary to sift through data and then simply report the results of the trial, without analyzing the trends and uncovering valuable information that is often hidden. In many cases, a biovisualization tool is a viable way to free up in-house staff to work on in-house development, research and infrastructure projects rather than having them sift through data.

The purpose of biovisualization platforms is not to perform a large amount of rigorous statistical analysis, but to offer different quantitative biovisualization techniques to explore the data and formulate hypotheses that can then be discussed and used to initiate further analysis — the purpose is to have a conversation with the data to better understand it, which makes the decision-making process more efficient. A company might wish to terminate a trial, depending on the results it is getting, or it may be inspired to further investigate a drug-specific biomarker.

IN SUMMARY

The importance of data management in clinical trials cannot be overstated. To cite a recent example: A pharmaceutical company found that trial subjects were experiencing adverse side effects. Yet, the alarm was not set off by the medical monitors, but by the data management team, emphasizing the increasing role that advanced data analysis will play in making rapid, mission-critical decisions and bringing drugs to market.

> *"The commonality between science and art is in trying to see profoundly — to develop strategies of seeing and showing." – Edward Tufte*

Biovisualization tools and their attendant interpretative services provide scientists and medical monitors a way to see their data more clearly — literally. The art of bringing data to life through simple visual displays may lead to profound observations revealed more quickly, ultimately adding speed and efficiency to the drug development process.

Author: Hermioni Zouridis, Senior Scientist, Scientific Operations