Paper PP20

Using Visual Dashboards to Improve Data Accuracy

Dan Smith, Fortrea, Leeds, UK

ABSTRACT

Data lies at the heart of any clinical trial, making its accuracy and reliability a primary objective for all stakeholders. Errors can arise in various ways, from typographical mistakes to incomplete data capture. Traditional approaches often rely on manual checks or reviews of derived datasets, which can increase the risk of human error or mask issues after analysis. This paper highlights the power of using simple visualizations based directly on source data to enable earlier and more accurate identification and correction of data errors. Visual dashboards aggregate and simplify large amounts of data, making it easier to spot anomalies and unexpected patterns. By implementing these visualizations, teams can improve inter-team communication, reduce the need for extensive data review, and minimize the risk of costly mistakes. Modern clinical trials have access to an increasing number of data visualization tools, making the adoption of a visual approach more accessible than ever.

INTRODUCTION

Clinical trials do not exist in a vacuum. They are defined by and respond to the requirements of the medical ecosystem they reside in. Trial development is required to evolve at a rapid pace owing to the ever-changing demands of new discoveries, novel treatment opportunities, and emerging diseases. A major symptom of these changes is that trials have been increasing in complexity for many years and continue to do so. A modern phase I study can look like a phase II or III study that was conducted even one or two decades ago. This complexity presents itself in numerous ways – number of sites across different countries, number of endpoints, number of eligibility criteria, number of procedures performed overall and per visit – all of which have been increasing for the past 15 years (Getz, et al., 2023). In response, the length of a trial has increased greatly and disproportionally. One review of over 16,000 phase I-III trials found that a 10% increase in trial complexity equated to a 33-36% increase in trial length (depending on phase) (Markey, et al., 2024). As trial length increases, so too do costs and overall workload. Though, pressure to mitigate against these increases remains high.

One pivotal way to counteract these increases is to ensure that accuracy of data is high and achieved early. A clinical trial fundamentally involves collecting and analyzing data, making data maintenance and validation essential to its process. If data is mishandled or incorrect, it threatens the validity and reliability of results and incurs timely and costly cycles of rectification and re-analysis, especially if errors are spotted late into the study. Though, with the number of data points for a given trial reaching into the hundreds of thousands (phase I) or millions (phase II/III) (Getz, et al., 2023), ensuring data accuracy is a difficult undertaking.

Data errors can arise in myriad ways (Table 1). Depending on the method of data entry, the rate of data errors can be as high as between 33 to 2,784 per 10,000 (Garza, et al., 2024). Similarly, one study found that 4.4% of over 40,000 eCRF pages that were entered into an EDC database contained some form of “data entry error” (Mitchel, et al., 2011).

Table 1

Data Error Examples	Description
Data entry errors	Errors that arise during initial data entry or transcription errors when copying the data from source to database – e.g. typographical errors.
Missing data	Entirely omitted fields or data points.
Inconsistency	Discrepancies in data entry between sites or ancillary data transfers.
Device or software errors	Equipment malfunctions or system data upload errors.
Protocol deviations	Data points that violate protocol-specified processes or time collection windows. Deviations may not always be flagged during data entry.

Monitoring and detecting these errors require numerous team members throughout the entirety of study conduct. Various strategies are implemented to organize and optimize the detection process, with site monitoring, data review cycles, and central monitoring being some of the most utilized (Knepper, et al., 2016). Traditional approaches are based on discrete batches of checks, starting with on-site source data review, continuing with edit checks and issue queries, and finalizing in SDTM/ADaM dataset and TFL review. This serial process suffers from limitations. (1) Errors can go unnoticed for weeks or months, increasing the risk of delayed timelines as corrective resolutions occur late into the study. (2) Multiple repeat on-site visits are labor-intensive and costly, where site monitoring already accounts for up to

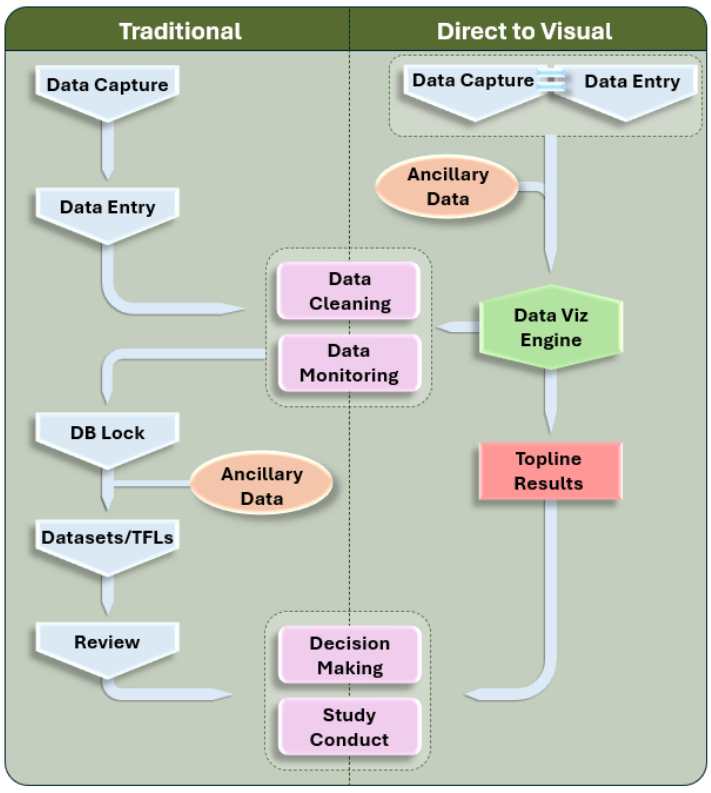
14% of total costs (Sertkaya, et al., 2016). (3) Segregated processes can lead to reduced cross-team communication and increase the risk of duplicating work (similar but separate reports and checks being performed). There is potential for these processes to be modified or replaced, and numerous pharmaceutical companies are currently undergoing this transition. Changes in and availability of new technology is one of the leading causes for biopharma companies to change their error detection strategies (Knepper, et al., 2016). The growing accessibility to multiple data visualization applications means that introducing visualizations as a novel detection strategy is an increasingly viable and sought after option.

The process outlined in this paper aims to retain the aspects of serial data review but consolidate them into one holistic, singular process. It outlines the use of direct-to-EDC data capture connected directly to a centralized hub of interactive data visualizations. These visualizations are tailored to the needs of those who are reviewing the data and are available live as the data is being entered. In doing so, the data review process is moved significantly earlier compared with the serial approach.

PROCESS OVERVIEW

The direct-to-visual process is designed to optimize four key areas: data cleaning, data monitoring, decision making, and study conduct (Figure 1). This paper primarily addresses the data cleaning and monitoring stages of the process. The first step involves utilizing direct data capture as the foundational source of data. This considerably reduces the number of errors from the outset by eliminating the data transcription step between data collection and entry. Data visualization tools are directly integrated with the database via various SQL queries, with automatic refresh intervals configured according to the task requirements or progress of data entry. During the early stages of data entry, the frequency of refresh is typically set between every 10 and 60 minutes. All visualizations are stored within a central dashboard so that internal team members and sponsors can efficiently access, interact, and navigate the various graphics in one location. Any identified data errors are documented, communicated to the relevant departments, and resolved.

Figure 1



VISUALISATION DESIGN PRINCIPLES

The purpose of a data graphic is simple: to communicate. Fundamentally, it translates quantitative data into insight by presenting proportionality, direction, and relationships. It reveals information otherwise unseen, converting an indigestible array of numbers into a palatable single concept. Getting the most out of a graphic requires incorporating specific features to ensure that the value of the information is optimal. It should present the data without distortion, incite investigation, encourage comparison of related pieces of information, and address a clear and specific need (Tufte, 1983).

Stasko (2014) outlines a descriptive equation to outline how valuable a particular graphic is:

$$V = T + I + E + C$$

V is the overall **value** of the graphic. **T** is the **time** required for the viewer to interpret the information. **I** is the level of **insight** it provides, whereby the most insightful graphic is one that reveals the most hidden information or promotes further investigation. **E** is the ability to convey the **essence** of the overall data. **C** is the amount of **confidence** that the graphic is representative of the original source.

The remainder of this paper is based heavily on these principles and sets to detail how they can be employed. In this context, the equation is modified to incorporate additional features deemed key to the overall process: **accessibility** and **simplicity**. Therefore, the modified equation is:

$$V = T + I + E + C + A + S$$

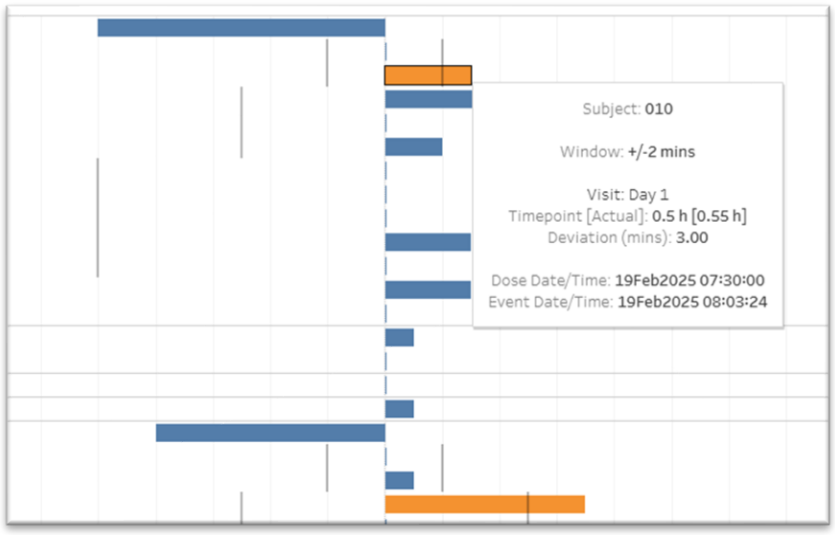
TIME

Time is an invaluable commodity. The longer that errors go unresolved the greater the risk of incurring costly resolutions and, when faced with a method of evaluating the data, the more time it takes to gain the necessary insights the more burdensome the workload. Humans are naturally adept at pattern recognition (Mattson, 2014), and visualizations leverage this strength. Outliers and trends can take more time to find or may be overlooked when viewed in datasets or spreadsheets alone.

A common requirement when collating trial data is to ensure that events are performed and samples are collected at the correct time. This involves cross-referencing multiple sample dates and times against respective dosing times. One approach that is typically employed is to use programmed edit checks to flag any outlying deviations. However, this can take time to create and validate and often only provides a single binary data point for each timepoint: whether there is a deviation or not. An alternative approach is to create a bar chart with the timepoints per participant plotted on the y-axis against the size and direction of the deviation on the x-axis (Figure 2). Protocol-defined windows are overlaid as reference lines where deviations beyond those lines are highlighted in a distinct color compared to those within the window. Hovering over the individual bars further reveals additional information related to the deviation that can be used for evaluating and communicating the possible error (Figure 3). Two key benefits are noted: (1) the use of color provides the ability to identify and navigate to outliers quicker than with a tabulator format; (2) organizing the data in this format allows for better detection of trends or patterns of deviations within or across participants.

Figure 2

Figure 3



In addition to isolating individual random errors, visualizations can effectively reveal systemic errors. Whilst all errors in data threaten the validity of results, systemic errors have been demonstrated to have a particularly deleterious impact on estimated between-group effects (Buyse, et al., 2017). For example, if a particular device consistently reports lower values or if a site shows consistent deviations in sample collection across a set of timepoints, this would be heavily pronounced in a graphic. Detecting these trends early allows for the necessary interventions – staff re-training, fixing faulty devices – to be implemented and potentially be triggered before the effects are permanent or costly to resolve.

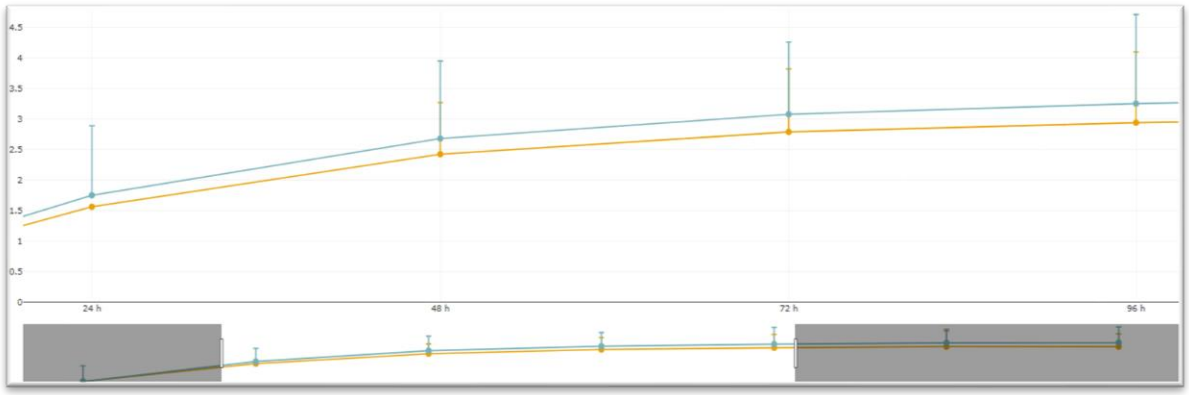
INSIGHT

An important step prior to creation of the visualizations for a study is to have upfront discussions with the individuals who will be reviewing the data. The type of data being presented partially dictates the best format to use. For example, when presenting adverse events by risk, dot plots and volcano plots are the preferred choice for evaluation (Qureshi, et al., 2022). These discussions are therefore critical to ensure that both domain expertise and data visualization best practices are driving the process.

Interactive visualizations should allow for dynamic adjustments of the data being presented. Features such as filters, buttons, and sliding scales grant control over specific portions of the overall graphic. Figure 4 displays a line graph of

drug concentration against time demonstrating the use of a sliding scale (presented underneath the x-axis) to adjust the timepoints being presented.

Figure 4



A user should be able to move up and down levels of detail where applicable. Gantt charts are a good example of this. Figure 5 shows a general overview of multiple events that occur during a study. Drug administration, reported adverse events, and concomitant medications taken per participant are plotted against study days. This creates a timeline for all events. Selecting any of the bars transforms the x-axis to be filtered only for the study day that the selected item occurred on (Figure 6). This highlights the ability to navigate up (global) and down (single study day) the level of details for a given study event.

Figure 5

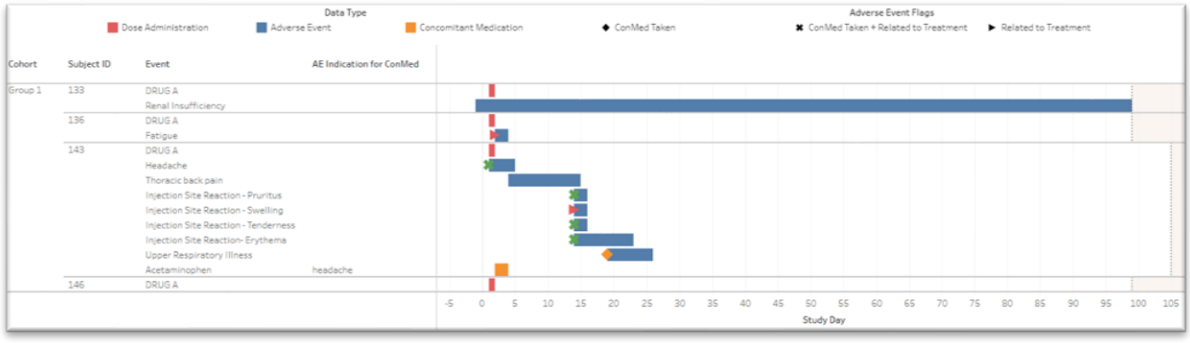
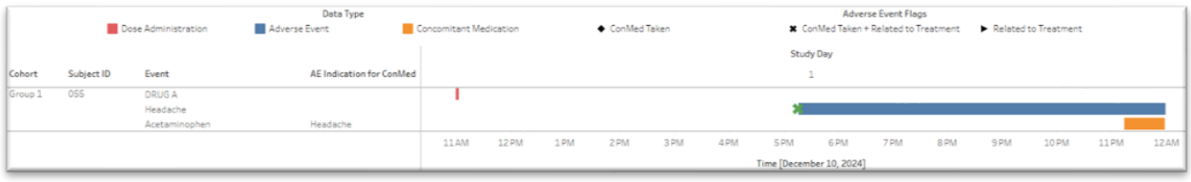


Figure 6



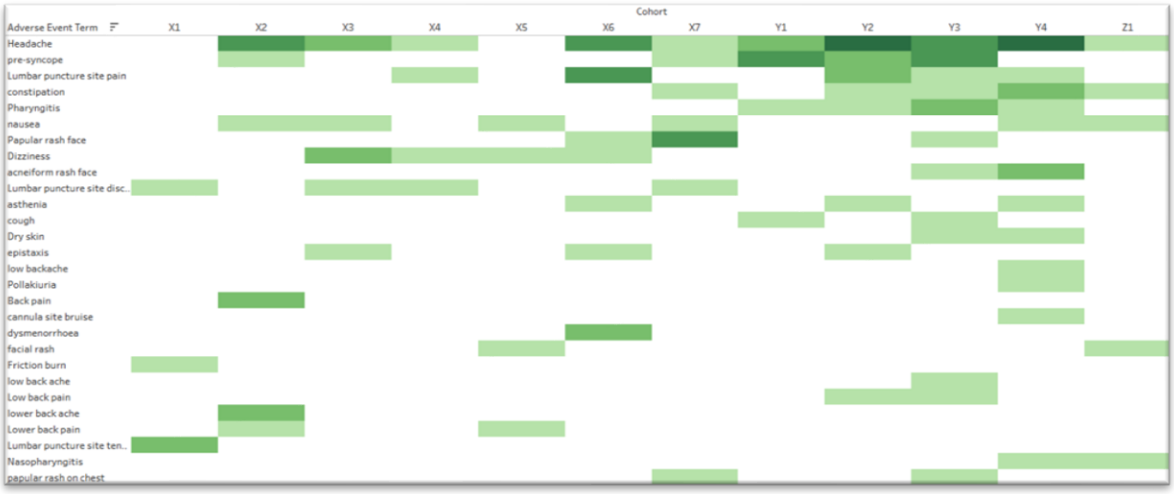
ESSENCE

The concept of essence within a single data visualization can be captured in two forms: (1) the **specific essence** of the intended data check and (2) the **general essence** of the study's data.

Extracting information from a graph is most accurate when the information being sought after is emphasized in the visualization. For example, it has been shown that estimations of probability are most accurate when the graphic emphasizes the shape of the probability distribution of the data (Heltne, et al., 2023). There must, therefore, be accurate and representative proportionality between the data and the visualization. Tufte (1983) describes this proportionality in terms of a Lie Factor, defined as the ratio of the size of the effect shown in the graphic vs the size of the effect of the data itself. A Lie Factor of 1 is therefore the optimal proportional representation. The sample collection date and time

deviation chart shown above (Figure 2) demonstrates these principles. Only one piece of information is required to be extracted, namely the size of the deviation. A bar chart is therefore chosen to best display the relative differences, plotted linearly so as not to distort the size of the effect, thereby preserving the specific essence visualization.

Figure 7



Heat maps are an effective way of representing the general essence of prevalence of study events. Figure 7 shows the distribution of adverse events stratified by cohort. Utilizing a color gradient allows for quick identification of common adverse events and the relative occurrence compared with less common events. It draws attention to, in this case, the cluster in top right of the graph which shows that the Y2, Y3, and Y4 cohorts had notably more events.

Use of cross-domain visualizations is another way to represent the general essence of a study. Traditionally used for project or task management, Gantt charts can be used in clinical settings to determine temporal dependencies or duplicate detection (Hirakata, et al., 2022). Here we highlight the power of Gantt charts for quick determinations of temporal dependency. The Gannt chart in Figure 5 conveys the overall distribution of events that occurred within a study, showing the emergence and cessation of adverse events and how they relate to dosing and non-treatment medications taken. To further highlight the dependencies, flags are included to indicate if the adverse event was treatment-related and/or if a medication was taken for it. These are common requirements for cross-domain data checks to ensure that data entered in one form (e.g. adverse event onset time and action taken) aligns with that entered in another form (e.g. concomitant medication onset and duration). The relationships between a participant's events are therefore presented holistically, reducing the need for reviewers to manually cross-reference separate data sources.

ACCESSIBILITY

Whilst visualizations offer considerable value, their creation and implementation require specialized software, technical expertise, and time investment that not all team members may possess. They should therefore be accessible to and interpretable by all stakeholders. To address this, a dedicated visualization team was established to develop and maintain the graphics, acting as a go-between for the source data and actionable insights. They are responsible for customizing the graphics to be role-dependent, so only the necessary data is being presented, and support the review or analysis being performed. By centralizing all visualizations into a single dashboard, gaps in technical knowledge are bridged for all teams, enabling all members to engage with the full breadth of data through a single interface. This hub is shared through a web browser, limiting additional access or software requirements.

CONFIDENCE

Confidence in a visualization refers to the extent to which the graphic is transparent and accurately represents the underlying data. By connecting the visualization tools directly to the source database, the number of steps between the original and transformed information is considerably reduced. Where feasible, calculated fields and elements during creation of the graphic are minimized to reduce the risk of incorrectly applied derivations. This has a secondary benefit in reducing the amount of validation of the output required. Each graphic undergoes peer review before being made available, focusing on validation of the database connections and calculations of derived variables and elements.

SIMPLICITY

Graphics are designed to direct the user's attention only to the intended insights. Features distracting the viewer from this should be avoided. In a review of visualizations of measurement uncertainty, one study found that plots that forewent overcrowding or presenting excessive information and tended towards simple, less detailed presentations were deemed the most optimal for evaluation (Heltne, et al., 2023). The authors noted that it is unknown whether this impact would be felt when the evaluators are trained clinicians, as they are trained to analyze the data for their investigations. Though, there still may be differential benefits of certain visualization formats within trained clinician cohorts. For example, quantile dot plots were shown to reduce interpretation errors in a cohort of clinicians when evaluating fictional clinical cases (Frans, et al., 2023). Once again, this outlines the importance of upfront discussions to determine the optimal formats and level of detail to present depending on team member's domains of expertise.

There are many ways that simplicity can be ensured (Evergreen & Metzner, 2013; Tufte, 1983): reducing "non-data ink" and "visual noise", defined as information that offers no aid to visual evaluation; removing unnecessary gridlines; avoiding excessive labelling, opting instead for interactive pop-ups; using colors and symbols, often in place of labels.

SUMMARY

Schneiderman (2003) outlines seven tasks required to be performed for a visual display to operate at its most optimal and deliver the best results (Table 2). The visualizations discussed here demonstrate and provide evidence for these tasks and how they can be implemented to enhance data monitoring processes.

Table 2

Task	Implementation
Overview	Storing all visualisations within a single dashboard ensures that the general overview of the suite of displays is readily accessible. Gantt charts (Figure 5) and heat maps (Figure 7) demonstrate ways to present the essence of specific domains of data or the overall study.
Zoom	Users should have the ability, where applicable, to zoom in and out of the visualisations so that specific portions of the data can be viewed depending on the query or task. Multiple-level Gantt charts (Figure 5 and Figure 6) and sliding scales (Figure 4) are examples of staggered and smooth zooming.
Filter	Sliders (Figure 4), drop-downs (Figure 2), and buttons are effective ways of allowing the user to control the data points being presented to focus only on the area of interest.
Details-on-demand	Once the user has navigated to the desired part of the visualisation, details should be available for the specific data points. Pop-ups when hovering over items (Figure 6) or links to separate data listings are effective approaches.
Relate	Cross-referencing of data is especially useful for evaluating relationships between data points. Timelines of events using Gantt charts (Figure 5) allow quick insight into the relationship between the actions and causes of distinct events.
History	A history of actions within the visualisations should be retained to allow reversion to prior views, an ability that is available in all presented displays.
Extract	All presented visualisations retain the ability to extract the source data and/or export a specific view of the graphic to various file types for use elsewhere.

CONCLUSION

Maintaining data accuracy is paramount for study validity, therefore making errors a burdensome cost. Early error detection thus becomes a pivotal process. To rise to the challenges of the changing landscape of drug development, companies must be swift and decisive about adopting novel strategies. By providing early, continuous, and intuitive insights, visual dashboards engender an interactive and collaborative monitoring process. Though, a one-size-fits-all approach should be avoided. Care must be taken to ensure the best form of visualizations are being presented for the given type of data. Specific attributes should be employed: simplicity rather than overburdening; insightfulness over generality. User-friendly visualization tools are more accessible than ever. The time to switch to a visual strategy is now.

REFERENCES

Buyse, M. et al., 2017. The impact of data errors on the outcome of randomized clinical trials. *Clinical Trials,* 14(5), pp. 499-506.

Evergreen, S. & Metzner, C., 2013. Design principles for data visualization in evaluation. *New directions for evaluation,* 2013(140), pp. 5-20.

Frans, N., Hummelen, B., Albers, C. J. & Paap, M. C., 2023. Visualizing Uncertainty to Promote Clinicians' Understanding of Measurement Error. *Assessment,* 30(8), pp. 2449-2460.

Garza, M. Y. et al., 2024. Error rates of data processing methods in clinical research: a systematic review and meta-analysis of manuscripts identified through PubMed. *International Journal of Medical Informatics,* p. 105749.

Getz, K., Smith, Z. & Kravet, M., 2023. Protocol design and performance benchmarks by phase and by oncology and rare disease subgroups. *Therapeutic Innovation & Regulatory Science,* pp. 49-56.

Heltne, A. et al., 2023. A systematic review of measurement uncertainty visualizations in the context of standardized assessments. *Scandinavian Journal of Psychology,* 64(5), pp. 595-608.

Hirakata, V. N., Oppermann, M. L. R., Genro, V. K. & Reichelt, A. J., 2022. Exploring the Gantt chart as a tool to highlight double report in case series published during the first wave of the COVID-19 pandemic. *Systematic Reviews,* 11(1), p. 155.

Knepper, D. et al., 2016. Detecting data quality issues in clinical trials: current practices and recommendations. *Therapeutic Innovation & Regulatory Science,* 50(1), pp. 15-21.

Markey, N. et al., 2024. Clinical trials are becoming more complex: a machine learning analysis of data from over 16,000 trials. *Scientific Reports,* Volume 14, p. 3514.

Mattson, M. P., 2014. Superior pattern processing is the essence of the evolved human brain. *Frontiers in neuroscience,* Volume 8, p. 265.

Mitchel, J. T. et al., 2011. Evaluation of data entry errors and data changes to an electronic data capture clinical trial database. *Drug information journal,* Volume 45, pp. 421-430.

Qureshi, R. et al., 2022. Comparing the value of data visualization methods for communicating harms in clinical trials. *Epidemiologic reviews,* 44(1), pp. 55-66.

Sertkaya, A., Wong, H.-H., Jessup, A. & Beleche, T., 2016. Key cost drivers of pharmaceutical clinical trials in the United States. *Clinical Trials,* 13(2), pp. 117-126.

Shneiderman, B., 2003. The eyes have it: A task by data type taxonomy for information visualizations. In: *The craft of information visualization.* s.l.:Elsevier, pp. 364-371.

Stasko, J., 2014. *Value-driven evaluation of visualizations.* s.l., Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization, pp. 46-53.

Tufte, E. R., 1983. *The visual display of quantitative information.* 2 ed. s.l.:Graphics Press Cheshire, CT.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Dan Smith
Company: Fortrea
Address: Drapers Yard, Marshall St, Holbeck, Leeds LS11 9EH
Email: daniel.smith@fortrea.com
Website: www.fortrea.com