Paper DV06

Straight to the Source: Interactive Visual Dashboards for Data Cleaning

Dan Smith, Fortrea, Leeds, UK

ABSTRACT

Data lies at the heart of any clinical trial, making its accuracy a primary objective. Errors can be costly, consume valuable time for rectification, and compromise integrity. Traditional approaches often rely on manual checks and reviews of derived datasets, increasing the risk of human error and delaying when reviews can occur. This paper highlights the power of using interactive visualisations based directly on source data to enable earlier and more accurate error detection. Upfront discussions with relevant stakeholders ensure that the visualisations are tailored to each department, allowing domain expertise to drive the process. These graphics are integrated into a centralized dashboard, enabling teams to collaborate in real time. Modern clinical trials have access to a suite of visualisation tools, making the adoption of a visual approach more accessible than ever.

BACKGROUND

Exemplified by the commonly used phrase that “data is the new oil”, the value of data is as high as it ever has been and continues to grow. This is especially true within the pharmaceutical industry, where the average clinical trial collects hundreds of thousands (Phase I) or millions (Phase II/III) of data points (Getz, Smith, & Kravet, 2023). Depending on phase, the estimated cost of a clinical trial is around $4 - $20 million (Wong, et al., 2014) or, stated otherwise, up to $6990 per participant (Speich, et al., 2018). Whilst costs are spread across numerous facets and components of the trial, according to the U.S. Department of Health and Human Services, data-related components (here grouped as Data Management, Source Data Verification (SDV), and Site Monitoring) of a Phase I trial account for 27% (~$574K per trial) of the overall costs (Wong, et al., 2014). A significant contributor to the costs associated with these data-related components is the identification and resolution of data errors. 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).

Data accuracy is therefore a prerequisite for a successful trial, but another key aspect is time. Reducing the overall time between data capture and access to data insights is a core requirement for all stakeholders. The median length of time between a patent application for a novel drug compound and its availability for prescription in US and Canada is 11.8 years (Lexchin, 2021). This equates to an average of 9.7 years of clinical development time (Brown, Wobst, Kapoor, Kenna, & Southall, 2022). To compound this, trials themselves are becoming longer and more complex. Between 2009 and 2020, there has been an upward trend of notable protocol design variables, including total number of endpoints (+27.1%), total procedures performed (+67.3%), and total investigative sites (+33.0%) (Getz, Smith, & Kravet, 2023). Similarly, between 2000 and 2005, the average development time for pharmaceutical companies increased by three percent (Getz K. A., 2006). However, the same study found that a handful of companies were able to reduce overall development time by up to 20%, demonstrating that implementing the right strategies has clear positive outcomes.

This paper outlines a strategy that utilises interactive visualisations to improve the data cleaning and monitoring capabilities. It demonstrates how these visualisations can be connected directly to the source data and updated in near real-time, moving the data review process significantly earlier compared with traditional approaches.

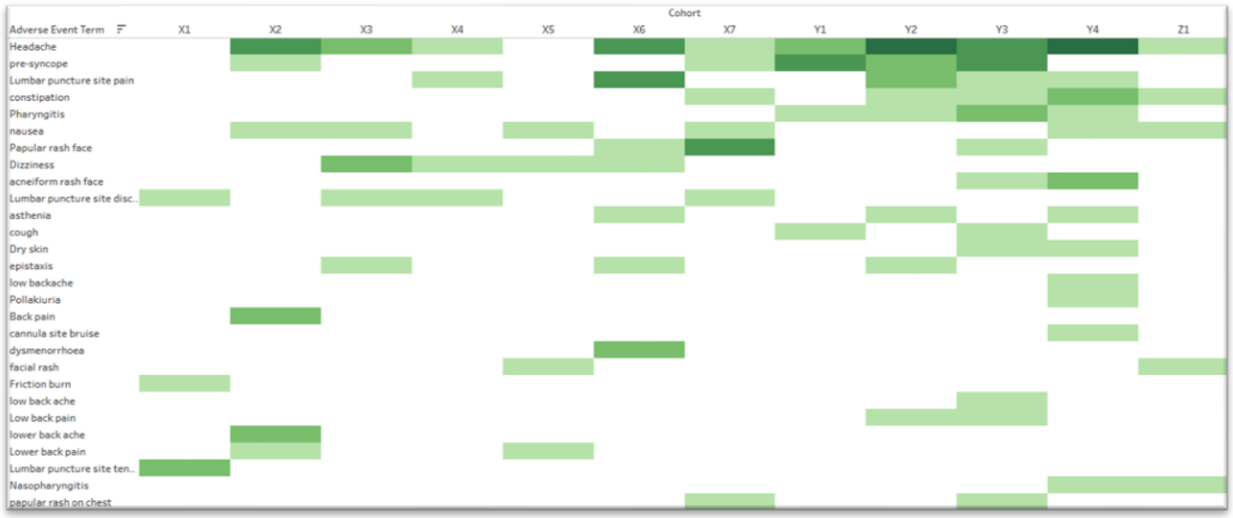
THE POWER OF VISUALISATIONS

Traditional approaches to data error detection rely on discrete rounds of reviews and checks starting with on-site source data review, continuing with edit checks and issue queries, and finalising with multiple rounds of SDTM/ADaM dataset and TFL reviews. This process flow is hindered by relying on manual checks throughout, slowed by the numerous procedural and programming steps required between deliverables, and generates only limited, static reports. Data errors can therefore go unnoticed for weeks, increasing the risk of delayed timelines as corrective resolutions occur later into the study.

Many pharmaceutical companies are therefore looking for alternative and novel approaches. One of the leading causes for biopharma companies to change their error detection strategies is changes in and availability of new technology (Knepper, et al., 2016). A modern biopharma company now has access to a wide array of programming environments, ranging from no-/low-code analytic tools (such as Spotfire, Tableau, and Power Bi) to open-source languages (such as R and Python). These tools offer comprehensive capabilities to create a suite of customisable and interactive graphics.

Visualisations are able to transform an overwhelming amount of data into an accessible and coherent form. They can be used to provide a general essence of the underlying data, providing fast and accessible top-level insights which can then be used to guide further investigations or inquiries elsewhere in the data or at more specific levels. The heat map in Figure 1 presents the distribution of adverse events across cohorts, demonstrating how the viewer's attention can quickly be drawn to the predominant cluster of adverse events within cohorts Y2, Y3, and Y4.

Figure 1

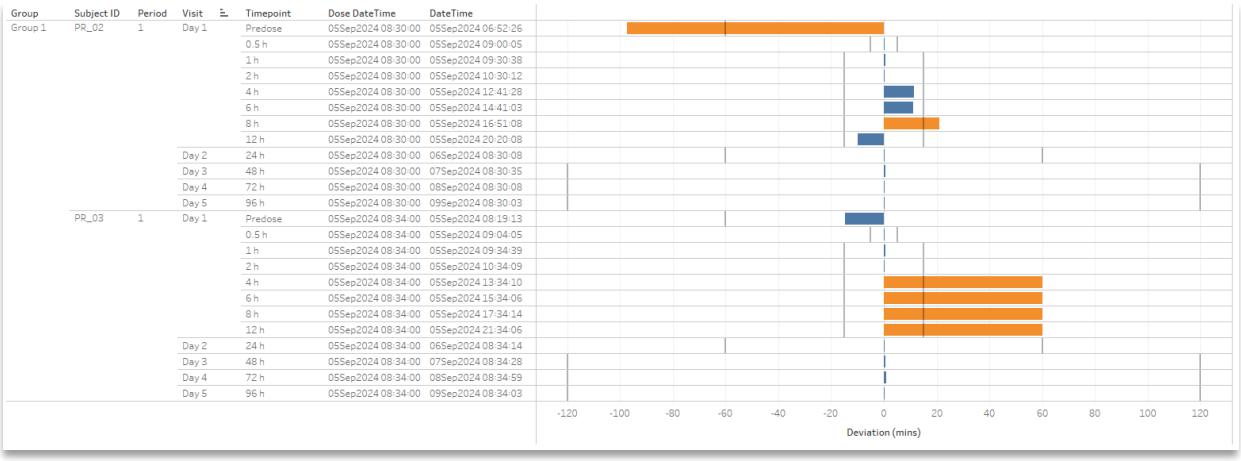


Humans are naturally adept at pattern recognition (Mattson, 2014), therefore translating data into graphical formats better orients reviewers to identify patterns and trends. Figure 2 and Figure 3 show the same data for sample collection date/time deviations being presented in tabular vs graphical formats, respectively. It highlights how anomalies are more easily identifiable when presented as a colour coded graphic compared with a standard grid of numbers with yes/no responses. Not only are the anomalies themselves more visible but the size of the deviation is presented in a clearer format. Rather than relying on slow comparison of individual values within a table, the user is able to determine the same information at a greater speed.

Figure 2

Group	Subject ID	Period	Visit	Timepoint	Category	Triplicate	Dose DateTime	DateTime	Full Window	Actual Time (hours)	Deviation (mins)	Out of Window
Group 1	PR_02	1	Day 1	Predose	PK		05/09/2024 08:30	05/09/2024 06:52	>-60 mins	-1.626111111	-97.56666667	Yes
Group 1	PR_02	1	Day 1	0.5 h	PK		05/09/2024 08:30	05/09/2024 09:00	+/-5 mins	0.501388889	0.083333333	No
Group 1	PR_02	1	Day 1	1 h	PK		05/09/2024 08:30	05/09/2024 09:30	+/-15 mins	1.010555556	0.633333333	No
Group 1	PR_02	1	Day 1	2 h	PK		05/09/2024 08:30	05/09/2024 10:30	+/-15 mins	2.003333333	0.2	No
Group 1	PR_02	1	Day 1	4 h	PK		05/09/2024 08:30	05/09/2024 12:41	+/-15 mins	4.191111111	11.46666667	No
Group 1	PR_02	1	Day 1	6 h	PK		05/09/2024 08:30	05/09/2024 14:41	+/-15 mins	6.184166667	11.05	No
Group 1	PR_02	1	Day 1	8 h	PK		05/09/2024 08:30	05/09/2024 16:51	+/-15 mins	8.352222222	21.13333333	Yes
Group 1	PR_02	1	Day 1	12 h	PK		05/09/2024 08:30	05/09/2024 20:20	+/-15 mins	11.83555556	-9.866666667	No
Group 1	PR_02	1	Day 2	24 h	PK		05/09/2024 08:30	06/09/2024 08:30	+/-60 mins	24.00222222	0.133333333	No
Group 1	PR_02	1	Day 3	48 h	PK		05/09/2024 08:30	07/09/2024 08:30	+/-120 mins	48.00972222	0.583333333	No
Group 1	PR_02	1	Day 4	72 h	PK		05/09/2024 08:30	08/09/2024 08:30	+/-120 mins	72.00222222	0.133333333	No
Group 1	PR_02	1	Day 5	96 h	PK		05/09/2024 08:30	09/09/2024 08:30	+/-120 mins	96.00083333	0.05	No
Group 1	PR_03	1	Day 1	Predose	PK		05/09/2024 08:34	05/09/2024 08:19	>-60 mins	-0.246388889	-14.78333333	No
Group 1	PR_03	1	Day 1	0.5 h	PK		05/09/2024 08:34	05/09/2024 09:04	+/-5 mins	0.501388889	0.083333333	No
Group 1	PR_03	1	Day 1	1 h	PK		05/09/2024 08:34	05/09/2024 09:34	+/-15 mins	1.010833333	0.65	No
Group 1	PR_03	1	Day 1	2 h	PK		05/09/2024 08:34	05/09/2024 10:34	+/-15 mins	2.0025	0.15	No
Group 1	PR_03	1	Day 1	4 h	PK		05/09/2024 08:34	05/09/2024 13:34	+/-15 mins	5.002777778	60.16666667	Yes
Group 1	PR_03	1	Day 1	6 h	PK		05/09/2024 08:34	05/09/2024 15:34	+/-15 mins	7.001666667	60.1	Yes
Group 1	PR_03	1	Day 1	8 h	PK		05/09/2024 08:34	05/09/2024 17:34	+/-15 mins	9.003888889	60.23333333	Yes
Group 1	PR_03	1	Day 1	12 h	PK		05/09/2024 08:34	05/09/2024 21:34	+/-15 mins	13.00166667	60.1	Yes
Group 1	PR_03	1	Day 2	24 h	PK		05/09/2024 08:34	06/09/2024 08:34	+/-60 mins	24.00388889	0.233333333	No
Group 1	PR_03	1	Day 3	48 h	PK		05/09/2024 08:34	07/09/2024 08:34	+/-120 mins	48.00777778	0.466666667	No
Group 1	PR_03	1	Day 4	72 h	PK		05/09/2024 08:34	08/09/2024 08:34	+/-120 mins	72.01638889	0.983333333	No
Group 1	PR_03	1	Day 5	96 h	PK		05/09/2024 08:34	09/09/2024 08:34	+/-120 mins	96.00083333	0.05	No

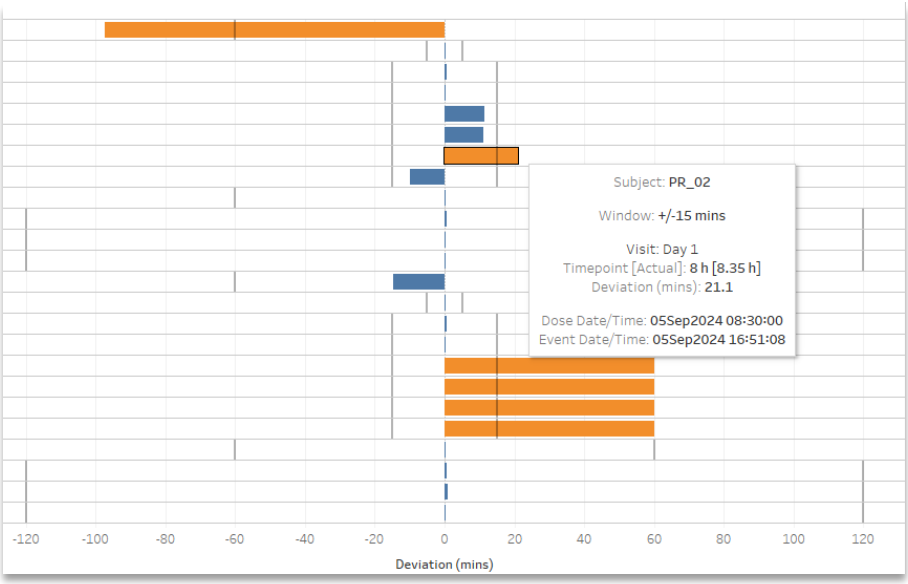
Figure 3



Group	Subject ID	Period	Visit	Timepoint	Dose DateTime	DateTime
Group 1	PR_02	1	Day 1	Predose	05Sep2024 08:30:00	05Sep2024 06:52:26
				0.5 h	05Sep2024 08:30:00	05Sep2024 09:00:05
				1 h	05Sep2024 08:30:00	05Sep2024 09:30:38
				2 h	05Sep2024 08:30:00	05Sep2024 10:30:12
				4 h	05Sep2024 08:30:00	05Sep2024 12:41:28
				6 h	05Sep2024 08:30:00	05Sep2024 14:41:03
				8 h	05Sep2024 08:30:00	05Sep2024 16:51:08
				12 h	05Sep2024 08:30:00	05Sep2024 20:20:08
			Day 2	24 h	05Sep2024 08:30:00	06Sep2024 08:30:08
			Day 3	48 h	05Sep2024 08:30:00	07Sep2024 08:30:35
			Day 4	72 h	05Sep2024 08:30:00	08Sep2024 08:30:08
			Day 5	96 h	05Sep2024 08:30:00	09Sep2024 08:30:03
	PR_03	1	Day 1	Predose	05Sep2024 08:34:00	05Sep2024 08:19:13
				0.5 h	05Sep2024 08:34:00	05Sep2024 09:04:05
				1 h	05Sep2024 08:34:00	05Sep2024 09:34:39
				2 h	05Sep2024 08:34:00	05Sep2024 10:34:09
				4 h	05Sep2024 08:34:00	05Sep2024 13:34:10
				6 h	05Sep2024 08:34:00	05Sep2024 15:34:06
				8 h	05Sep2024 08:34:00	05Sep2024 17:34:14
				12 h	05Sep2024 08:34:00	05Sep2024 21:34:06
			Day 2	24 h	05Sep2024 08:34:00	06Sep2024 08:34:14
			Day 3	48 h	05Sep2024 08:34:00	07Sep2024 08:34:28
			Day 4	72 h	05Sep2024 08:34:00	08Sep2024 08:34:59
			Day 5	96 h	05Sep2024 08:34:00	09Sep2024 08:34:03

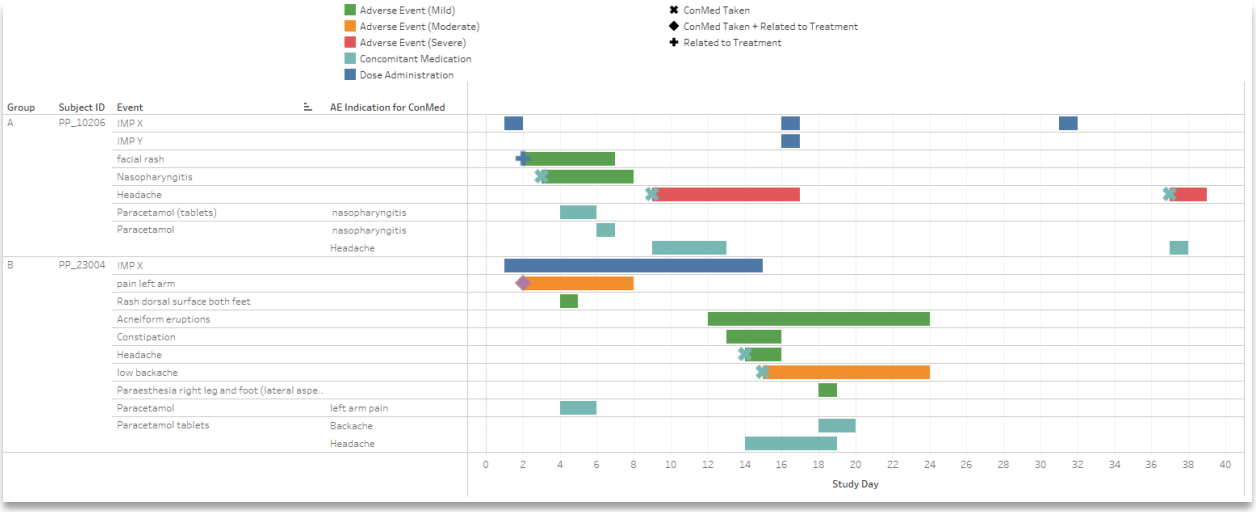
One of the drawbacks of traditional TFL-based approaches is the reliance on the static outputs. By transforming the graphic into a dynamic visualisation, the user is able to interact directly with the output. This is achieved using elements such as filters, action buttons, sliding scales, and hover-overs. Figure 4 demonstrates how, once the user has identified a data point of interest, hover-overs are available to gain additional insight and information.

Figure 4



Many graphics presented in standard submission packages represent summaries of single-domain data. However, it is common to need to compare different domains of data against one another. For example, temporal dependencies between study events need to be checked and validated to ensure that the dates and times of the events are logical. The Gantt chart in Figure 5 presents the distribution of adverse events, concomitant medications, and study drug administration across the duration of the study. By displaying these study events in a timeline, the emergence, cessation, and interdependence of the events are summarised and clearly presented. Additional levels of detail are included in the visualisation: colours are used to distinguish the type of event and their severity, where applicable; and flags are overlayed to demarcate the adverse events that are treatment-related and/or required medications to be taken. With this style of presentation, the user is able to holistically assess multiple data types and how they relate to one another.

Figure 5



KNOWING YOUR AUDIENCE

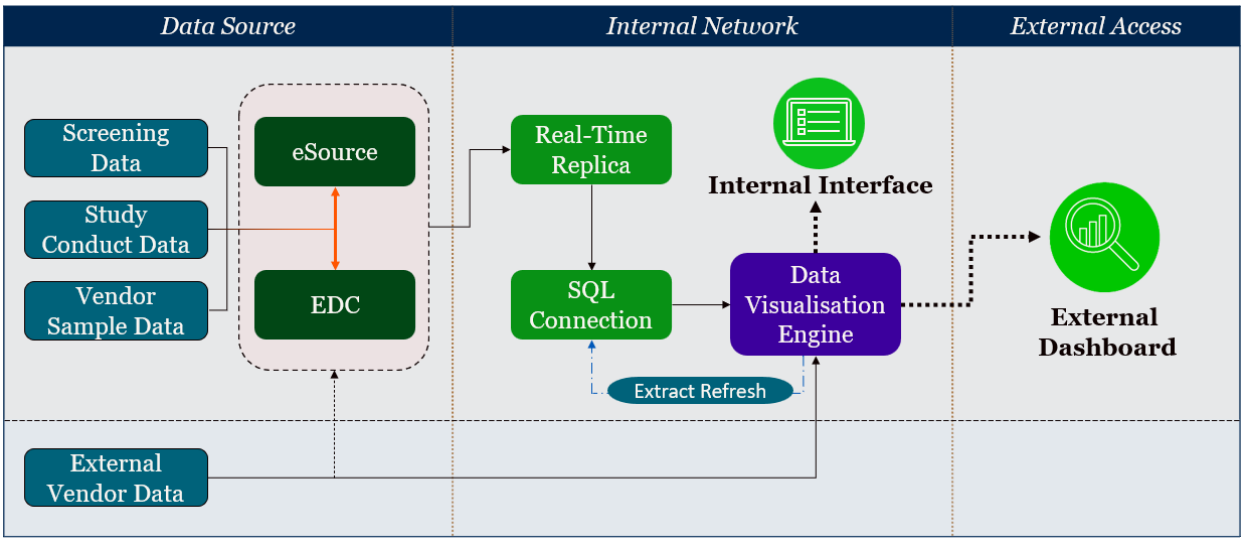
Whilst visualisations provide a powerful tool for data review, not all graphics are suitable for all types of investigations. The type of data being presented should partly guide the decision on which visualisations to present. For example, one study found that when presenting adverse events by risk, dot plots and volcano plots are the preferred choice for evaluation (Qureshi, et al., 2022). An important step before creating visualisations for a study is therefore to have upfront discussions with the various team members who will be evaluating the data. This allows their domain expertise to drive the overall process to further improve the impact the visualisations have on monitoring and evaluations.

Once these visualisations have been decided upon and are developed for a project, they can be housed within in a centralised hub or dashboard. By granting multiple users to a single accessible area, teams are able to improve their intra- & inter-team communication as all members can view the data in the same format and at the same time. This provides a second layer of efficiency as, not only are possible anomalies being detected sooner, but data queries can also be communicated more readily, thereby improving the efficiency of raising data queries.

PROCESS OVERVIEW

This process (outlined in Figure 6) aims to incorporate the two key aspects for data monitoring discussed above: (1) timely access to the data and (2) leveraging the power of visualisations for insights and evaluations.

Figure 6



I. eSource & EDC

- Utilising the ClinSpark® eSource system, data at the site is collected electronically, rather than on paper, and captured directly and automatically into the EDC. This is stored primarily in the Master database. Any updates or amendments to the data are logged automatically and available to be viewed immediately to all those with access to the system.
- No transcription stage is required, as the EDC data is itself the source data, thereby greatly reducing the burden of Source Data Verification. Source Data Review is performed to validate and cross-check the data that has been collected.
- By cutting out the transcription stage between eSource and EDC, the overall rate of data errors is significantly reduced from the outset by removing the potential for human error inherent in the transcription process. It also truncates the duration of the data collection process, minimising delays to downstream processes.
- **External Vendor Data**
    - Not all data is collected at internally hosted sites and is instead transferred from an external vendor. This data can also be loaded directly into the data visualisation tools to provide early insight using the same principles for presenting database information.

II. Real-time Replica

- Read Replicas are created as direct copies of the Master database. The replicas are updated continuously and instantaneously, with only a 20-millisecond delay between Master and Replica.
- These synchronised replicas are used as the source for downstream connections as they are MySQL compatible databases, providing the capability to generate live connections between the visualisation engine and the database.

III. SQL Connection

- A connection between the Replica database (client) and the data visualisation tool (server) is established using MySQL. This session fetches the necessary datasets and variables within the data for use in the visualisations.
- The connection is stored locally as an extract and embedded within the visualisation tool's environment or package. This extract is a static snapshot of the dataset at the time of extraction, providing a usable dataset for creating visualisations without depending on live queries which can be burdensome and slow to respond.
- A schedule is then configured that updates the extract at intervals that are required for the investigations, ranging from minutes, to hours, to days. The frequency of the refreshes is dependent on the purpose of the visualisation. For example, data points related to key protocol endpoint will be scheduled to refresh in near-real-time.

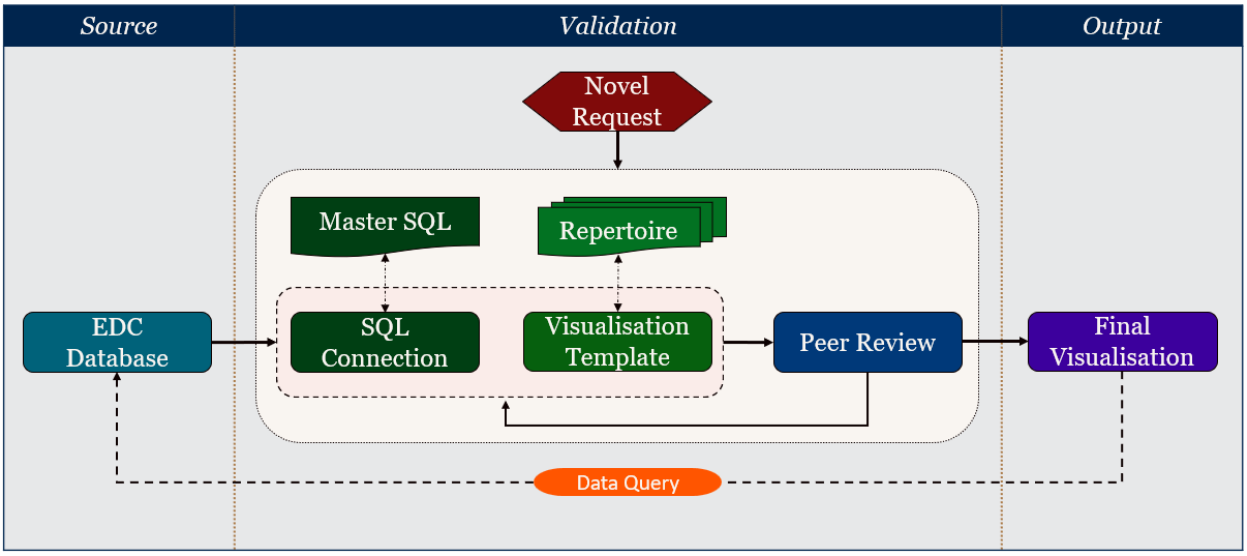
IV. Data Visualisation Engine

- Using the pulled database extracts, the visualisations are created using the preferred tool. Initially, before data entry has begun or has limited data points, the visualisations are generated using the UAT (User Acceptance Testing) database, allowing the developer to create the foundations for the visualisation and verify the formatting and appearance of the output. As more data becomes available, the visualisations are updated and finalised using the production database at which point the variable mappings and study-specific data is checked to be accurate and is being presented as expected.
- Combining upfront discussions with reviewers and monitors about which visualisations are required, the use of validated templates (discussed in the preceding section), and use of UAT databases ensures that overall programming time is shifted earlier into the study timeline so that all visualisations are live and available on the day of the first dose.
- Internal users – such as clinical operations staff, data managers, and statisticians – are granted access to the individual outputs within a dedicated project workspace. Whereas external users – sponsors and collaborators – are provided with secure access to a single dashboard hosted on an external server. The dashboard is designed in a tabbed structure whereby each tab navigates to a different data type and/or visualisation type.

VALIDATION

Between initial data collection and the final visualisations, there are several steps that require validation to ensure that all data being presented is accurate, representative, and comprehensive (outlined in Figure 7).

Figure 7



Database & Replica

In this workflow, given that the eSource and EDC pipeline is combined into a single unified entity, validation of the data collection itself is built into the system. Whilst this approach does not entirely eliminate the possibility of data errors, as the data entry itself can still be collected incorrectly, it does guarantee that the visualisations will accurately reflect the source data as it was collected. There are no unseen alterations or transformations made between collection and output. The implementation of visualisations, as well as other internal data checks, allows for identification of any data collection errors thus providing a feedback loop of validation of the source data.

SQL Connection

To ensure that the connection between the database and the visualisation tool is retrieving the correct and full data, the SQL connection itself must be validated. This process begins with the creation of a Master SQL script, which is written to pull a set of aliases, forms, and variables that encompasses all the relevant data necessary for the intended roster of visualisations. Demographics, adverse events, study drug administration, concomitant medications, vital signs, and ECGs are an inexhaustive example of the data types included in the Master SQL. Validation of the Master SQL is achieved by systematically comparing the data extracted via the SQL connection to the original database extract for the same aliases and forms. This comparison is performed via a dedicated in-house programming application that compares every data point, checking for any discrepancies, omissions, or mismatches and generates a report of any findings. A clean report signifies that the SQL has successfully pulled all anticipated data and is therefore validated.

Once the Master SQL has been validated, each visualisation is then created using an SQL script based on a subset of the Master. This subset will include only the required data types for the intended graphic. For example, as in Figure 5, only data related to adverse events, concomitant medications, and dosing will be included thereby removing any unnecessary configurations which could otherwise be burdensome to processing speeds. If a visualisation is requested for a novel data type or if the database is built using a different structure from the Master SQL, the comparison validation process is repeated for that situation. This iterative approach ensures flexibility over time as study or sponsor requirements evolve.

Visualisation Templates

To reduce the development time for each project, blank templates are used as the starting point for development. These blank templates are preloaded with the Master SQL so that the development programmer can then retain or remove any aspects as required so that only the relevant data is being pulled. All graphical features are also predefined, including formatting for colours, symbols, legends, aspect ratios, hover-overs, and filters, which will automatically adjust for the study data.

Each template also includes the majority of calculated fields and variable mappings required for the given visualisation – for example, timepoint formats and date/time assignments and imputations. By embedding these commonly used calculations within the template, repetitive programming is avoided across multiple projects. To determine a comprehensive template code, the visualisation is tested on numerous mock and prior studies that use the standard database build structures and formats. This ensures that the templates are well adapted to numerous possible study designs and data formats, thereby reducing the overall number of changes and inclusions that the development programmer may need to make to finalise the output. The peer review process is consequently streamlined as consistency is better maintained across different visualisations and studies and the possibility of errors or incorrect mappings are reduced. All templates are monitored, reviewed, and improved as time goes on, creating an ever-evolving repertoire of available visualisations for teams to choose from.

Peer Review

The final validation step is peer review of the final visualisation. Final checks are performed by an independent programmer who assesses the derived variables, data presentation, and formatting. All derived variables are evaluated for their calculations and transformations. Presentation of data, including all graphical elements (such as colours, symbols, and legends), is reviewed to ensure that all the relevant formats are accounted for and display as expected. Interactive elements (such as filters, hover-overs, and clickable actions) are also tested to ensure a seamless user experience. Importantly, study-specific mappings are evaluated against the annotated casebook, which includes all naming conventions for the database forms, aliases, and variables. This step ensures that all database items are being accounted for and assigned to the correct corresponding visualisation element. Any review comments are fed back to the development programmer and implemented until final, in the same manner as typical programming standards.

CONCLUSION

Accurate and timely data insights are fundamental to ensuring the efficiency of clinical trials. With the increasing availability of numerous visualisation tools, teams have the capability to monitor and evaluate study data in novel ways. A visual approach affords significant advantages over the traditional data review processes commonly adopted. To further extend the power of visualisations, by connecting them directly to source-data, real-time evaluations can be facilitated to greatly improve data error detection strategies. Engaging in upfront discussions with reviewers puts domain expertise at the centre of the process and ensures that a responsive and adaptive suite of visualisations are constantly available.

This paper opens with the idea that “data is the new oil”; however, an alternative perspective from David McCandless (2010), that “data is the new soil”, is equally fitting in this context – it requires care to cultivate. The direct-to-visual approach highlights that, with the correct processes in place and implementation of robust validation, data can be made available in near real-time in a format that is highly accessible to all.

REFERENCES

Brown, D. G., Wobst, H. J., Kapoor, A., Kenna, L. A., & Southall, N. (2022). Clinical development times for innovative drugs. *Nat Rev Drug Discov, 21*(11), 793-794.

Garza, M. Y., Williams, T., Ounpraseuth, S., Hu, Z., Lee, J., Snowden, J., . . . Young, L. W. (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*, 105749.

Getz, K. A. (2006). Hitching a Ride with the Speed Demons of Drug Development. *Applied Clinical Trials, 15*(12), 22.

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*, 49-56.

Knepper, D., Fenske, C., Nadolny, P., Bedding, A., Gribkova, E., Polzer, J., . . . Lawton, A. (2016). Detecting data quality issues in clinical trials: current practices and recommendations. *Therapeutic Innovation & Regulatory Science, 50*(1), 15-21.

Lexchin, J. (2021). Time to market for drugs approved in Canada between 2014 and 2018: an observational study. *BMJ Open, 11*(7), e047557.
Mattson, M. P. (2014). Superior pattern processing is the essence of the evolved human brain. *Frontiers in neuroscience, 8*, 265.
McCandless, D. (2010). The beauty of data visualization [Video]. TED Conferences. Retrieved from https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization
Mitchel, J. T., Kim, Y. J., Choi, J., Park, G., Cappi, S., Horn, D., . . . D'Agostino Jr, R. B. (2011). Evaluation of data entry errors and data changes to an electronic data capture clinical trial database. *Drug information journal, 45*, 421-430.
Qureshi, R., Chen, X., Goerg, C., Mayo-Wilson, E., Dickinson, S., Golzarri-Arroyo, L., . . . McAdams DeMarco, M. (2022). Comparing the value of data visualization methods for communicating harms in clinical trials. *Epidemiologic reviews, 44*(1), 55-66.
Speich, B., %A von Niederhäusern, B., Schur, N., Hemkens, L. G., Fürst, T., Bhatnagar, N., . . . Pauli-Magnus, C. (2018). Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data. *Journal of clinical epidemiology, 96*, 1-11.
Wong, H.-H., Jessup, A., Sertkaya, A., Birkenbach, A., Berlind, A., & Eyraud, J. (2014). Examination of clinical trial costs and barriers for drug development final. *Office of the Assistant Secretary for Planning and Evaluation, US Department of Health & Human Services*, 1-92.

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