

Janssen DATA4YOU platform approach for clinical data review

Paul Fioole, Janssen R&D, Beerse, Belgium

Erik Lemmens, Janssen R&D, Beerse, Belgium

Pieter Pluymers, Janssen R&D, Beerse, Belgium

ABSTRACT

The paper discusses our initiative to transform the clinical data review environment through the DATA4YOU program. In response to the challenges posed by increasing demand and the complexity of clinical data, Janssen undertook a phased approach to modernize their data review processes and tools. The paper highlights the key objectives of the initiative, which start from the needs of the different main user groups: Clinical Data Managers, Medical Reviewers, Central Monitoring/Risk Management, and Clinical Programmers. It describes the gaps in the existing processes, underscoring the importance of near real-time data accessibility, automation, and improved user experience. Then, it highlights how the collaborative and user-centric approach was adopted in developing the DATA4YOU application.

The authors outline the strategic approach taken based on a cloud-based best-of-breed solution. They detail the technical aspects of a solution for an integrated data review environment consisting of a cloud-based centralized DataMart and web application integrated with visualization tooling. The paper underscores the innovative features of the system, such as versioning capabilities and complex data tracking, enabling users to review advanced clinical reports efficiently.

INTRODUCTION

Janssen is launching a new collaborative clinical data review solution as part of a multiyear and staged program to renew our clinical patient data environments.

At this point, we are going through the first phase, which targets modernizing the clinical data review environment for medical review, risk management, central monitoring, and data management.

By aligning review processes across different teams and utilizing shared tools and processes, Janssen aims to significantly reduce rework across teams and enable us to keep pace with the increasing demand.

The program is set up in a modular way and is part of the overall architectural renewal of our clinical environment. The scope of the current phase is reviewing tooling and processes for our teams and the underlying data layer and data sources, focused on making data available in a single DataMart. This platform will be extended gradually for other consumptions and other feeds of new data sources.

APPROACH TOWARD A FUTURE-ORIENTED ENVIRONMENT

THE CURRENT CHALLENGES

The DATA4YOU program drives the process of transforming clinical data from different data sources into information [DATA] and making this data available to Key Stakeholders promptly to engage in a collaborative review of the data [4YOU].

There is a growing pressure on processing and reviewing more complex and voluminous clinical data, as well as an increasing need to get the patient data earlier and flawlessly.

Our current processes have evolved to cope with this, but we must keep up with the increasing demand involving new methodologies and automation.

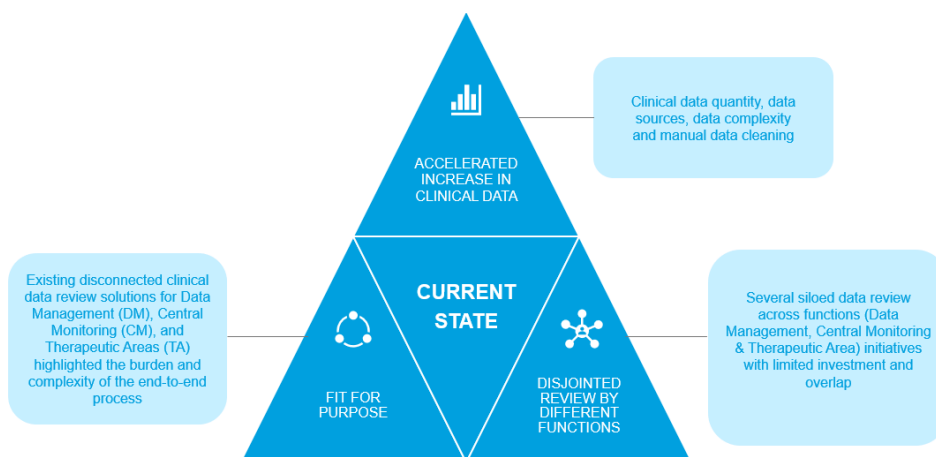


Figure 1 - current challenges

These are supported by platforms for storing and processing clinical data, reporting, and visualization tools but have limited integration. On top of these leading platforms, more manual (e.g., Excel-based) processes are used to complete the work. This results in an environment that is suboptimal and less fit for use, difficult to learn, and requires several process deviations, which is becoming increasingly burdensome, to keep up with the growing demand as described.

As different departments have different requirements, they have created historically siloed point solutions to cope with specific review tasks, leading to a more disjointed review. This limits the ease of collaboration, which is becoming a requirement in this increasingly complex data environment.

TOWARDS DATA4YOU

We started with a Strategic Alignment Exercise and completed an assessment to identify pain points and define desired functionality. We defined three major gap categories in order of priority:

1. **Gaps in managing and automating the workflow for E2E data review.** Information and feedback related to the data quality review must be shareable across different teams instead of being handled in other solutions and manually exchanged via e-mail as executed today. This called for a standard solution that can be shared amongst our business partners to facilitate query management with our Clinical Sites and suppliers. Therefore, we required more automation in a shared workflow solution that starts from a common integrated process orchestrating the individual actions in the distinct groups working on dataflow and data review.
2. **Suboptimal and separate solutions for data review and visualizations.** To maximize user satisfaction, the aim is to design and implement a future-proof, next-generation data review and visualization environment. This innovative environment must be capable of being shared across different teams, replacing the largely separate solutions we have today.
3. **A common data layer for clinical data that is cost-efficient and scalable.** This will ensure a sole source of the truth with accessibility for different requirements. Currently, data is being copied into multiple distinct locations for consumption and data files are duplicated many times and are scattered over various locations. There is no good way of centrally managing the different versions. LSAF is using SAS programming language, while we want to be open to other more modern programming languages. On top of that, our next-generation clinical data store should be able to support “future-oriented” tools for dataflow automation, transformation, and optimization. This will increase the speed of getting access to “near-real-time” data as a targeted outcome.

Based on these gaps, we decided to work on three major elements to achieve the desired outcomes:

- We defined and agreed on a strategy for the implementation of a unified **platform** that will enable us to have near real-time data accessibility, data processing, and visualization that is future-proof and can respond to the evolving needs of the organization.
- We defined a clear and unified **end-to-end data review process** that is fully endorsed across the different departments, so we can both get the desired collaboration across teams and the enablement through technology and automation.
- We defined a **cross-functional system** for data review workflow that leverages the process and is equipped with a user-friendly interface to effectively execute the process and engage the different teams using it.

PLATFORM SOLUTION APPROACH

During the strategic alignment phase, we evaluated options for technical implementation.

A decision was taken to opt for a best-of-breed approach to develop our future environment. This way, we can enhance our environment based on what is available in the market and maintain/develop where appropriate. This will also allow us to have a more modular approach so we can better prioritize urgent needs (like the collaborative review environment). Another strategic choice was to build on components and solutions that allow for a more open structure in terms of technology.

On the technology layer, we are following the cloud-first paradigm, which has become the mainstream. Many sources and downstream applications we are integrating with are already implemented on JNJ's cloud VPC environment, which supports GxP applications, so we decided to base DATA4YOU on this platform. We are looking at leveraging cloud-based solutions and services that allow for scaling to performance, which has been a significant issue for some of our current in-house platforms.

DATA4YOU – AN INTEGRATED USER-CENTRIC APPROACH

In every transformational project, people and processes are essential. During the DATA4YOU alignment, we spent a lot of effort understanding and documenting the needs of the people who will need to reap the benefits of the improvements we are targeting. Based on this information, we developed the targeted end-to-end process and mapped this towards the gradual implementation of our platform.

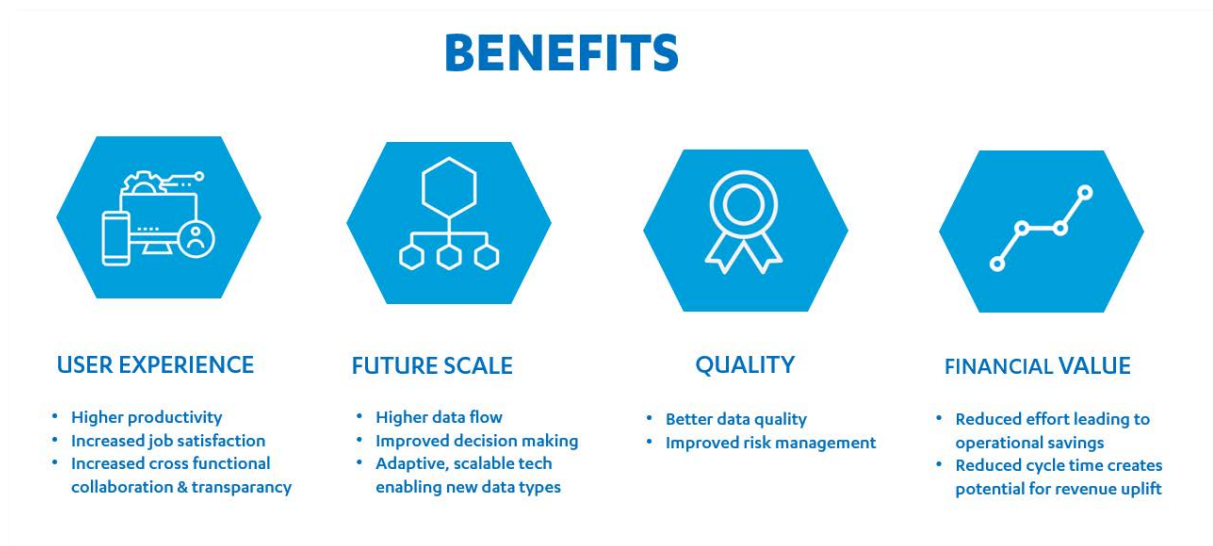


Figure 2 - Main benefits from a business perspective

One of the items that resonates very strongly is that all of our users insist on an improved user experience for executing their jobs. The current effort leads to extra work, more copies of the same data, and initiates more communication needs between different teams using dissimilar tools to enable more cross-functional collaboration and transparency. We see an ever-increasing demand for our data flows and reviews, with shortened and more critical timelines that our teams need to cope with. This also calls for implementing novel approaches in handling complex and new data types and to come to better decision-making.

Our users have a demand for improved quality of the data, which goes back to more consistent and undoubtedly timely data. Bringing this to the next level will allow us to manage our risks better.

Lastly, with the amount, volume, and complexity of our trials, we also see costs rising significantly. Implementing the DATA4YOU end-to-end process in a novel way with more automation, integration, and collaboration will allow us to keep these costs under control.

PERSONAE IN DATA4YOU

We defined four primary users that will start using DATA4YOU in the first phase. These are listed below. Benefits for each of them might be different depending on their roles, as expanded upon below. Today, they are also using disparate solutions that, in the new world, need to be unified or at least be connected. Below is an overview of the main feedback we did get from the different personae concerning the expectations for DATA4YOU



Figure 3 - Personas involved in clinical data review

CLINICAL DATA MANAGERS

Their focus is on the completeness and quality of the data so that others can quickly derive insights regarding the trial. They focus on the details and want to ensure they can dedicate the necessary time to the critical issues. To achieve this, they are asking for easy and fast access to the data at different levels. Near real-time data is essential. At the same time, the ask is to get an intuitive and easy-to-use interface, as the current disparate systems ask for too much back-and-forth, and some require specific skills.

They also expect tools to easily follow up on the quality of the data, especially data considered critical for the study analysis and safety follow-up of patients, with the necessary communication to different parties, including clinical site personnel.

MEDICAL REVIEWERS

Their focus is on ensuring the products are safe for the patient, as well as on the quality of the data and preventing protocol deviations. Given the close follow-up needed, they want to get updated quickly on the latest information available. While they might be using the same review environment, their focus is more targeted towards combining data and organizing it in real time so they can get more insights and recommendations. The system should also be able to point the attention to critical events or trends by enabling more analytical use

cases. Lastly, the collaboration features should allow them to understand if someone has already reviewed and cleaned the data that they are looking at and understand if the site has already been asked for information or data review.

CENTRAL MONITORING/RISK MANAGEMENT

They care most about finding trends and potential risks in the data to ensure the success of trials. The organization depends on them to identify issues quickly, so fast access to the data is paramount. Hence, they also require quick data access so they can immediately start identifying trends. They need an easy data request process and data available in near-real-time. Dashboards should be quickly configured.

CLINICAL PROGRAMMERS

In our trials, clinical programmers are implementing and configuring the dataflow, including the setup of the review listings and visualizations. Today, there is a multitude of solutions to configure with dedicated tools; they are asking for a more end-to-end view based on integrated tools and easy-to-use configurations based on a modern UI and leveraging modern techniques like low/no coding solutions. There is also a significant demand for a study-oriented monitoring dashboard that allows the teams to keep track of the status of the data flow and the quality of the data.

END-END PROCESS FOR DATA CLEANING AND REVIEW

A revised “To-Be” end-to-end process for data review and cleaning was developed as part of our process initiative. As an outcome, we now have the design of a revised functional end-to-end process in its “Future To-Be” state. A detailed process map for an optimized study setup and study conduct is being defined and is serving as the basis to define automation and build out our workflow solution as part of DATA4YOU. This creation was executed as a collaborative exercise involving the groups from the different personae to ensure all requirements were integrated.

The review process includes the approach for:

- Critical items list
- Integrated Data Review Plan
- Approach to Query and Discrepancy Management (including automation strategy)

During the DATA4YOU requirements gathering, this process was further detailed into individual Use Cases that could be linked to the implementation requirements for the solution development. From here, we worked out the process as we wanted to see it implemented in the system.

INTEGRATED DATA REVIEW – A COLLABORATION-ORIENTED APPROACH

DATA4YOU is developed and implemented gradually; in the first phase, we focused on the integrated data review. This is where we determined that we had most of the pain points and gains were the highest. We targeted to build out the initial platform, ensuring we can service the main personae user groups as described above.

The implementation is following our end-end process for data cleaning and review. The starting point for implementation in the platform is based on the integrated review plan. This plan is maintained in the system and contains the review requirements, which describe the goal of a planned review brought together for the different personae. As this is a collaborative exercise, this ensures that we are working across the different teams in an integrated way.

Review requirements are linked to their corresponding review methods. You can read these as all technical implementations to verify and describe the data so reviewers can make meaningful conclusions. These can be visualizations (like all kinds of graphs), listings – which can be simple to very complex, AI/ML-supported reporting, or even automated checks, e.g., for structure review.

The actual work to execute the plan is enabled through tasks, which are distributed across the different personae who are involved in the review. These tasks can be assigned by the system, and users can assign them to themselves to allow for flexible execution of the plan.

As there is integration across the different teams, we enable a real collaborative environment that spans all people involved in the execution of the review of a trial.

You can refer to the associated presentation slides to review screenshots of how our users currently can execute the tasks. When they find data issues during review, they can create inquiries. These can be assigned to other teams or be used to create queries towards the site EDC system. This can all be done based on the DATA4YOU platform and for all teams together this time.

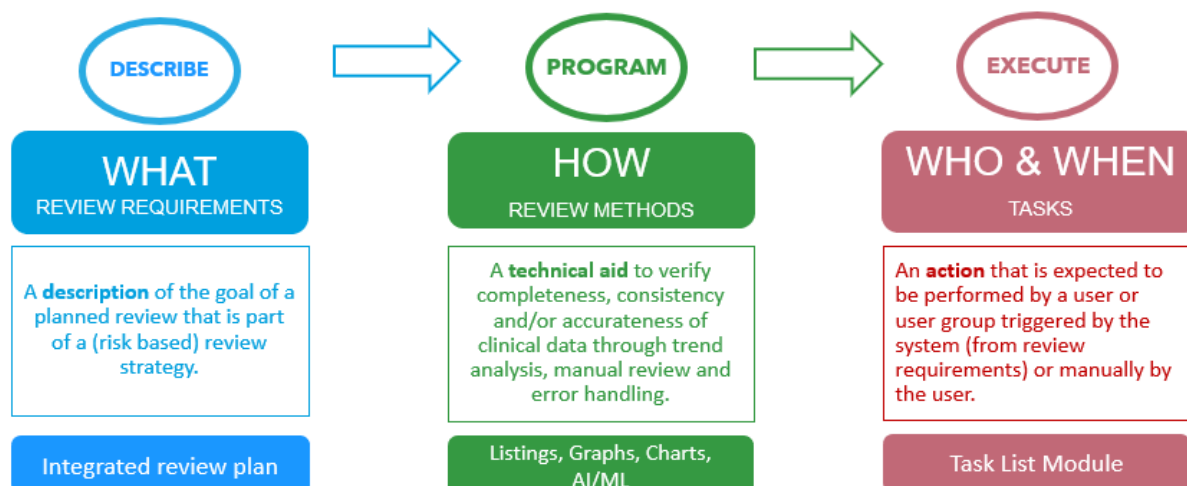


Figure 4 - Integrated review from plan to action

INNOVATIVE INTEGRATED DATA REVIEW ENVIRONMENT

The first implementation as part of DATA4YOU is our data review environment. This is based on the following main components: a Datamart based on a Databricks Lakehouse, a web application that integrates the different end-user interactions within the system, supporting cloud-based infrastructure to manage the processes, workflow, and a relational database to support the review processes, and a Tibco Spotfire backend to render visualizations that are integrated into the web application.

The data used by Spotfire, and the web app resides in the Lakehouse (see below). This separation of data and compute allows us to better scale and ensure that we can adequately manage the performance of the composite system.

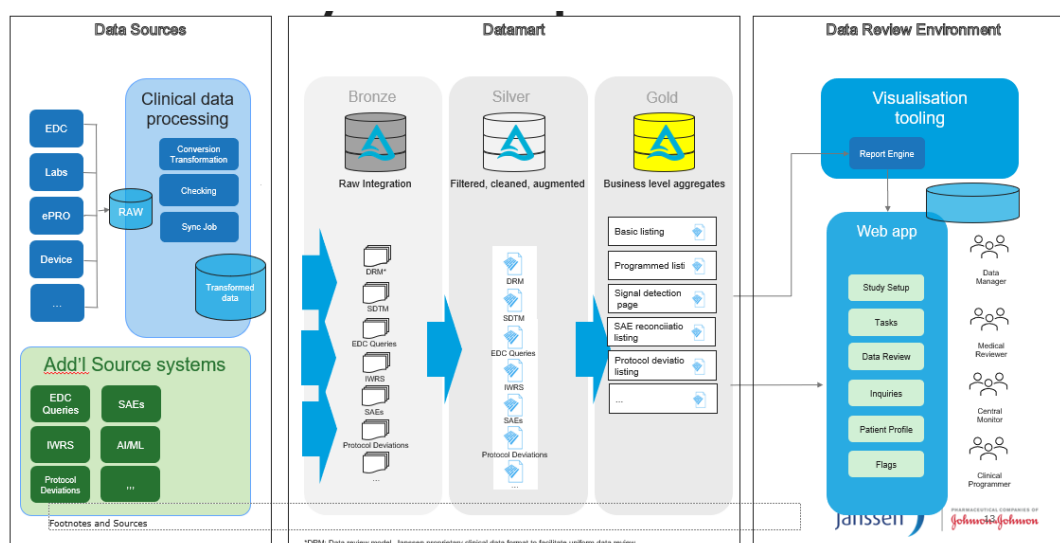


Figure 5 - current system solution structure

DATA SOURCES

Data is retrieved from different sources, so we can support diverse data that can be composed through the Databricks Datamart. The main source is our clinical patient data, retrieved from the clinical data repository where we store and process clinical data from raw to reporting data. Still, there are currently several other sources like protocol deviations or Serious Adverse Events from our safety system, and other AI/ML-based reporting that can be integrated into the review.

DATAMART

Our goal with the DataMart setup was to ensure we do have a flexible common data source that can scale to the needs of feeding the data review processes, can easily be integrated with our existing data sources, and has the necessary enabling tooling to drive our data pipelines for the complex review requirements we do have on the data review side. Together with our solutions provider, we decided to use a data layer that applies the so-called Lakehouse concept.

Such LAKEHOUSE features the flexibility and cost-efficiency of a data lake for storing data with the contextual and high-speed querying capabilities of a data warehouse.

For our case, we see a data lake that can work very well with ingested files on the incoming side, while towards the outgoing side, we present data more as if it were a data warehouse. The data layer is based on Delta Lake, which is open source and accessible from other services, e.g., visualization platforms or possibly other compute nodes such as external statistical computing servers.

Data pipelines typically follow the so-called medallion architecture for dataflow, which is generally based on three steps:

- Incoming data is fed into the lake house in the bronze layer, which receives the data and keeps all versions in single tables. This will allow for later reference and historical searches. Here, we feed our data from different sources.
- Next, the data is cleaned and structured into the silver layer, where the versioned metadata of the individual records is maintained. This gives powerful information for review and tracking of history (so-called time travel), allowing to report on subsequent versions of the data. This enables us to bring the review reporting to a much more user-oriented level.
- Finally, in the gold layer, data is aggregated, filtered, and selected to provide targeted data tables for the Data Review Environment. This gold layer is the actual access to the DataMart, set up for best performance, accessing only needed data. This means a great step forward versus our current situation, where we keep copies in different systems. This is one of the reasons why our current application is slow, requiring expensive hardware upgrades.

Data tables are configured in the gold layer based on the review requirements per study. A central catalog is easing this process, ensuring we can maximally reuse table definitions across studies.

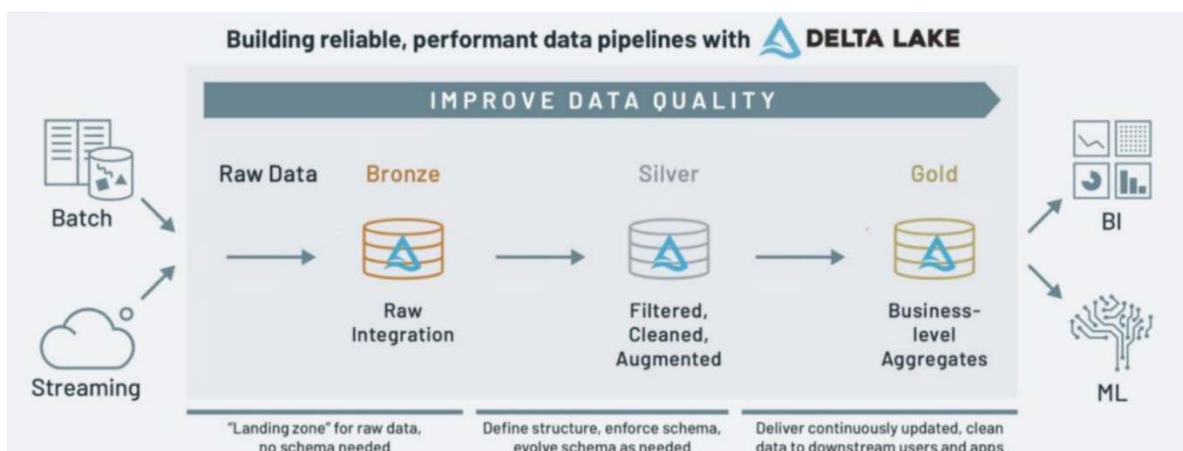


Figure 6 - Medallion architecture

Tables in Delta Lake are called delta tables and are based on Parquet files. This is an optimized column-oriented storage and can be optimized via built-in capabilities. The Delta table functionality also adds versioning capability, allowing for time travel, and ACID transactions for atomicity (any transaction will succeed in full or not at all). Incoming data transfers can also be rolled back, which is another significant requirement for us as data is highly dynamic, and inconsistency in incoming data might happen. These features are where the storage platform differentiates itself from regular data lakes.

WEB APPLICATION

The web application we developed with our solutions provider is at the heart of the review system. It is implemented in our VPC cloud together with the other components of the Lakehouse DataMart. Users are serviced by a web application that offers a user-friendly interface to all functions they need. A relational database is implemented to keep track of review statuses, inquiries, and queries so users have that information in combination with the review data at their fingertips.

BRINGING IT TOGETHER: TRACKING COMPLEX DATA REVIEW

The Delta Lake tracks individual changes between different versions of sourced data. This will help to ease the data review by our users, saving them time in doing their work and keeping track of review statuses on a per-record level in their review listings. Thanks to the Lakehouse architecture, we can implement this both on the base data (as it is delivered from the source systems) and on combined and complex data. Below is a simplified representation of a combined data record: a medical reviewer might have a review requirement that requires checking adverse events vs related concomitant medications for selected patients.

Updating the base tables will also trigger the complex table – which combines the AE and CM data– to be updated. Based on the basic listing versions, we can determine whether the complex listing version needs to increase and present the complex listing record as an updated record to the end user. Consequently, the user can select only the records that they need to (re-)review, which is a great gain of time.

Thanks to the data mart and the way we set it up in combination with the web, this allows for a powerful and performant tool to enable our users in their complex day-to-day review.

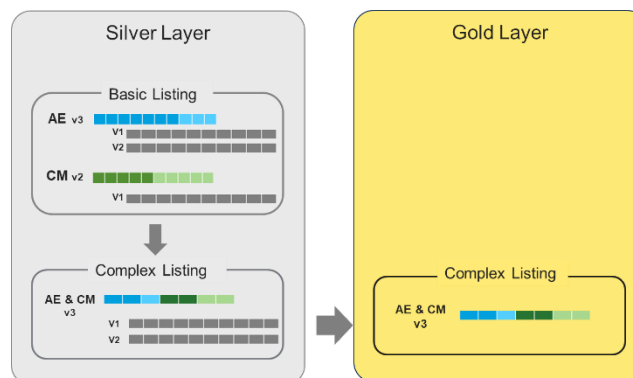


Figure 7 - complex data record tracking

CONCLUSION

In conclusion, the DATA4YOU initiative represents a significant step forward in the realm of clinical data management at Janssen by addressing the challenges posed by escalating data volumes and intricate review processes. The emphasis on a unified platform, real-time data accessibility, incorporating AI/ML, and user-centric design is resulting in the development of a sophisticated integrated data review environment.

The adoption of a best-of-breed approach, coupled with the utilization of cutting-edge technologies, underscores Janssen's commitment to innovation and flexibility. The initiative's collaborative orientation, involving diverse user groups such as Clinical Data Managers, Medical Reviewers, Central Monitoring/Risk Management, and Clinical Programmers, ensures that the developed solutions are tailored to meet the specific needs of each stakeholder. By enhancing data quality, improving efficiency, and promoting seamless collaboration, DATA4YOU not only resolves existing challenges but also positions Janssen well for dealing with future needs and opportunities in the evolving clinical data landscape.

CONTACT INFORMATION

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

Paul Fioole

Janssen R&D

Email: pfioole@its.jnj.com

Web: <https://www.linkedin.com/in/paul-fioole-985469/>

Erik Lemmens

Janssen R&D

Email: elemmens@its.jnj.com

Web: <https://www.linkedin.com/in/erik-lemmens-666a212/>

Pieter Pluymers

Janssen R&D

Email: ppluyme2@its.jnj.com

<https://www.linkedin.com/in/pieter-pluymers-099783a/>

Brand and product names are trademarks of their respective companies.