



PhUSE US Connect 2019

AB02

Maximizing the Use of Patient Data Warehouse via Visual Analytics

Sandy Lei, Janssen Research & Development, Titusville, US
Simson Alex, Johnson & Johnson, Spring House, US

ABSTRACT

Clinical trial data warehouse plays an important role in the drug development cycle when addressing questions on patient data by compounds, indications, therapeutic areas, or even companywide clinical trials. Common concepts such as patient enrollment rates, subject discontinuation reasons, signal detections from adverse events, patient diversity, Hy's law subjects by compound, to name a few, are large scale data extraction and analysis efforts when involving hundreds of studies. They demand intensive efforts by various compound teams.

The Janssen Patient Data Warehouse (PDW) was established in 2013 as a platform for hosting clinical trials data for disease area stronghold compounds and the associated metadata and documents. Within this platform, users are empowered with a document browsing tool - Trial Catalog, and data retrieval tools - Study Browser and Data-on-Demand. The extracted data can be imported into user's preferred analysis tools for downstream research and analysis.

PDW Visual Analytics platform contains end-to-end analysis applications that generate analysis results dynamically based on metadata selections of studies, datasets, variables, and variable values. Design and development of this effort started in late 2017. Using a commercially available visualization tool, we have enhanced our framework with the power of on-demand user-driven querying across hundreds of studies. Pre-existing analysis templates facilitate the automatic population of analysis results upon query initiation. This allows us to focus on creating business context specific analysis templates for addressing various questions as one-time efforts that are then dynamically populated for varying user selections, bringing significant value to business partners.

This dynamic visual analytics platform will be demonstrated with some PDW use cases. Data standard conformance will be presented via cross-study metadata comparisons. Benefits and challenges will be discussed.

INTRODUCTION

The Janssen Patient Data Warehouse (PDW) was established in 2013 as a technology platform enabling the exploration of patient data across Janssen clinical trials. As of today, the PDW database contains over 480 clinical trials data and metadata based on the original SAS dataset storage folders that support the Tables, Lists and Figures (TLF) for clinical study reports. The Trial Catalog application is a high-level document browsing tool on this platform. For extraction of clinical trials data, Study Browser and Data-on-Demand are the data retrieval tools provided. Data can be retrieved in two forms: As integrated datasets for the studies of interest, or as all study datasets for subjects selected by user defined criteria. The retrieved data are stored under user specific locations. Users may perform analysis using any tool of preference such as SAS BASE/STAT, R, Tableau, Spotfire, Qlik or others. The PDW Visual Analytics is a capability with a similar design that additionally includes analysis capabilities through a visualization tool. This capability brings the ability to design and build visualization and analytics rapidly, thereby allowing us to move towards maximizing the use of PDW in many clinical trial development processes.

The PDW platform is employed to support questions related to multiple studies, compounds, indications, therapeutic areas, and companywide clinical trial data. Even when multiple studies are being queried that can number in several hundreds, the goal is to facilitate the ease of use. Automated data integration will, however, generate meaningful analysis-ready datasets only if the selected cross-study data conforms to the same data standards. Since data is loaded into PDW as is from the original study results, conformation to data standards may not always be the case, leading to the need for ad hoc cleansing, and making a case for upfront cross-study harmonization and data unification efforts in the future. With increasing data content, another challenge is that of run time query performance.

BACKGROUND

In Janssen, clinical trial data that supports analysis and reporting for regulatory submissions, reside on various file servers. Data, programs and outputs are stored under study folders and subfolders. This allows for the sequential access of a handful of data files for analysis. For large scale cross-study integration, analysis and reporting, reliance on SAS programming is high and costly. In 2013, the data warehouse group in Janssen Research and Development collaborated with the J&J Information Technology organization to design and develop the Patient Data Warehouse platform with clinical trials data and metadata stored in a centralized relational database, which afforded us with the ability to quickly access data across studies and made way for the use of modern techniques in data extraction, integration and analyses. In late 2017, we embarked on designing and developing the PDW Visual Analytics platform by incorporating end to end processes involving data selection, integration, analysis and visualization.

DESIGN AND APPLICATION

Infrastructure: A three tier architecture is used to implement the PDW platform as shown in Figure 1. A relational database is used in the backend to persist the clinical trials data in a structured data model. Web and visualization tools described in the following paragraphs are the user interfaces that query and display data. A RESTful (Fielding, R) application programming interface layer forms the main part of the middle layer that can communicate between the database and the user interactions. It also communicates with a SAS server as needed.

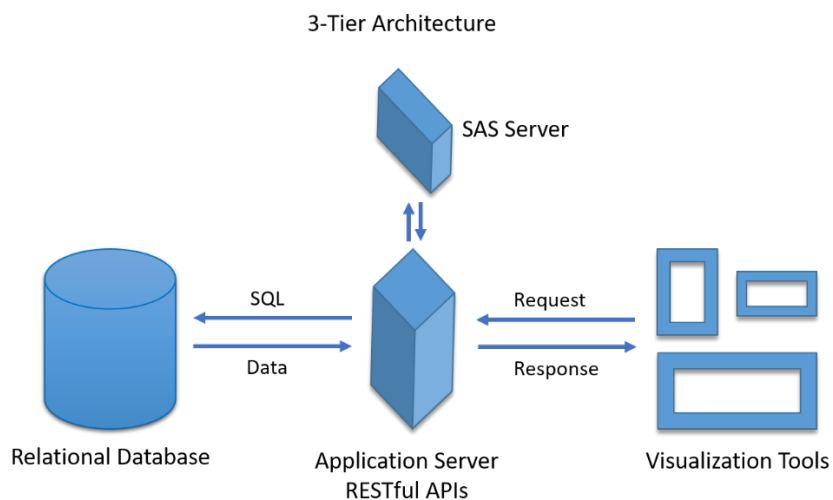


Figure 1. A simple 3 tier technology architecture is implemented and optimized for maintenance, performance and scalability

PDW Database Schema and Data Loading: The main data components of the PDW database include five types of metadata (study, dataset, variable, variable-value, parameter-value) and patient value data. These various data elements are stored in a normalized data model in a relational database. The loading of data into the database is an automated process that is triggered upon the identification and verification of data sources that are SAS datasets created from clinical studies. As of 2018, PDW contains patient data from over 480 studies. This corresponds to over 1200 reporting efforts (database lock milestones). It contains analysis, tabulation and raw data from over 40,000 SAS datasets. This further corresponds to over 1 billion SAS dataset rows and over 40 billion patient value data elements.

PDW Study Browser: The Study Browser tool, which provides new capabilities for automated data integration (also known as pooling or stacking), is designed to retrieve SAS data tables (in their original format) via metadata selections of studies, datasets, variables, and variable values. These selections drive the creation of the necessary SQL queries to be executed on the database. A SAS program is also generated that serves to execute the SQL queries and save the data under user specific locations. A batch job server (BJS) and a SAS 9.4 server are used in tandem to achieve the data extraction and data is generated in SAS, CSV or XPT formats as requested. Integrated datasets mean that a single data file is created for a domain. Business rules for the data structures were defined to facilitate the automated data integration, similar to the rules for ISS/ISE data pooling. Metadata selection is the focus in the use of Study Browser and hence the user is expected to have domain knowledge to achieve desired results. Open ended questions, however, require robust search and exploration capabilities.

PDW Data-on-Demand: Functionally similar to Study Browser, the Data-on-Demand tool provides additional user-

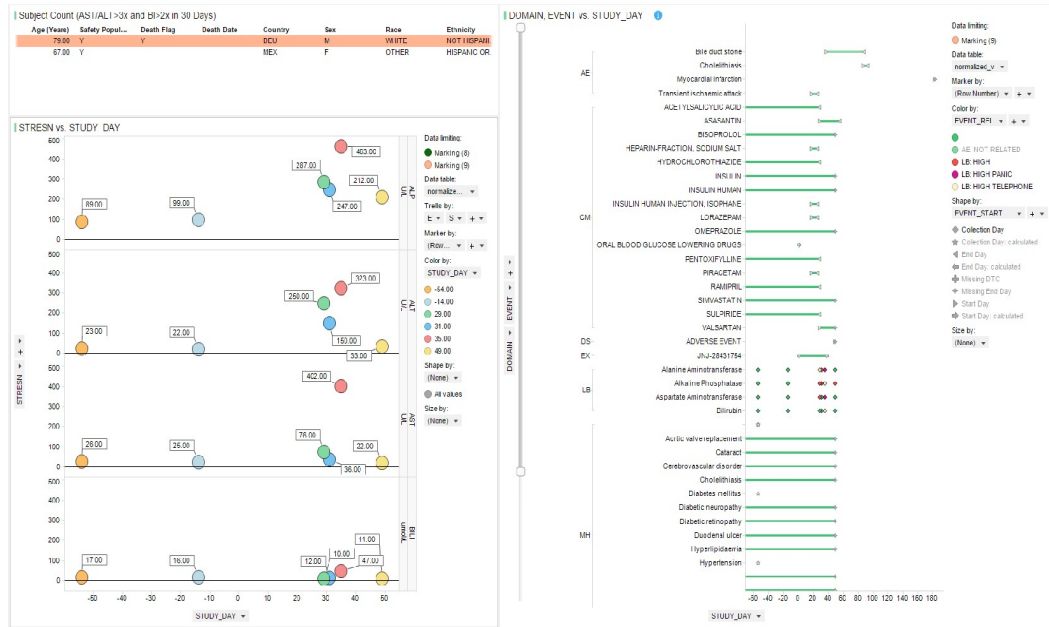


Figure 3. A patient profile visualization of a Hy's Law subject

PDW Use Case 2 (PDW Visual Analytics – flexibility in visualization layout): Provide demographic summary statistics and/or distributions for study data conforming to SDTM from 3 selected compounds

Figure 4 shows the visualization upon selection of 3 compounds in a bar chart on top-left, the correspondent data types (Analysis, Tabulation and Raw) in a tree map are shown on bottom-left, upon the selection of Tabulation data type (in red), all applicable reporting efforts in a table are shown on the right, 46 studies are selected (in green).

Figures 5 and 6 show the datasets and variables selection capabilities, respectively. Finally, Figure 7 shows the results of the analysis. Subject counts by country from 46 studies, age summary by race (box plot), race distribution (bar chart) and age distribution by gender (bar chart) are depicted. It is evident that even with the SDTM conformed datasets, data harmonization for race categories is needed.

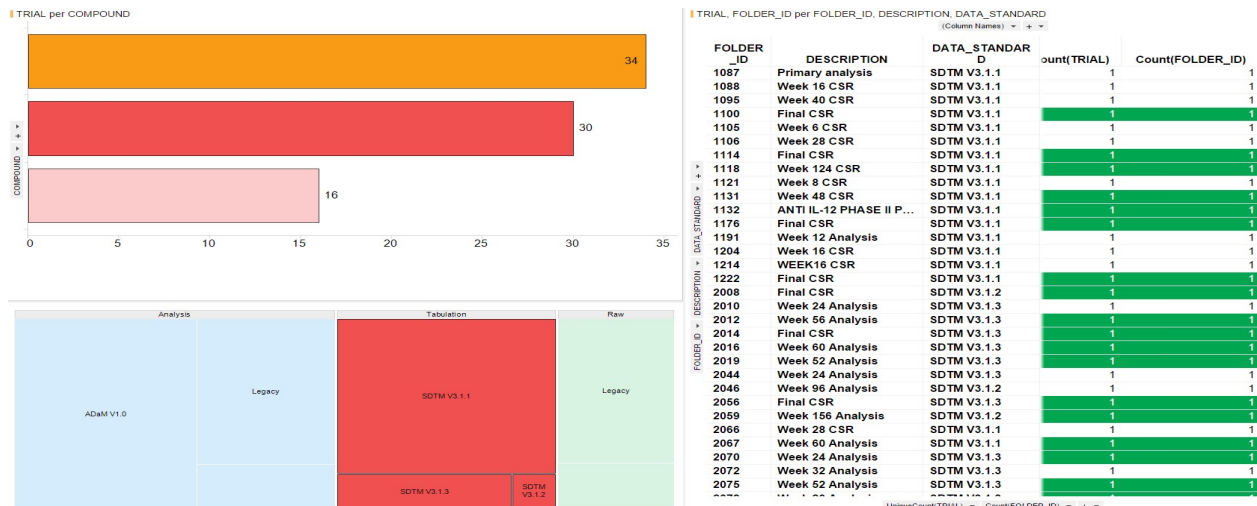


Figure 4. Selection of 3 compounds, Tabulation data type and applicable reporting efforts in 46 studies. Upon the study reporting efforts selection, all available datasets in the 46 selected studies are shown in Figure 5.

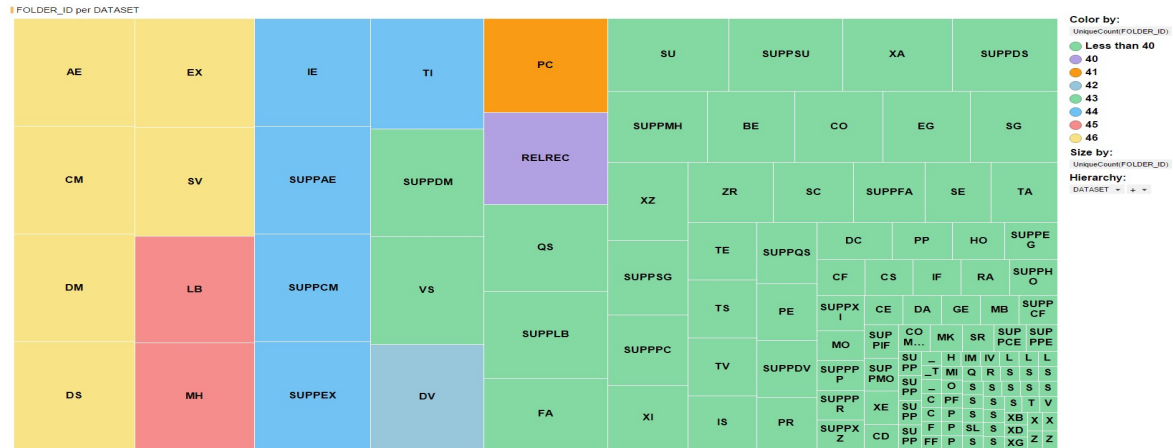


Figure 5. Dataset selection capability. Domains AE, CM, DM, DS, EX and SV shown in yellow are present in all 46 selected studies. Select data AE, CM and DM to view all available variables in Figure 6.

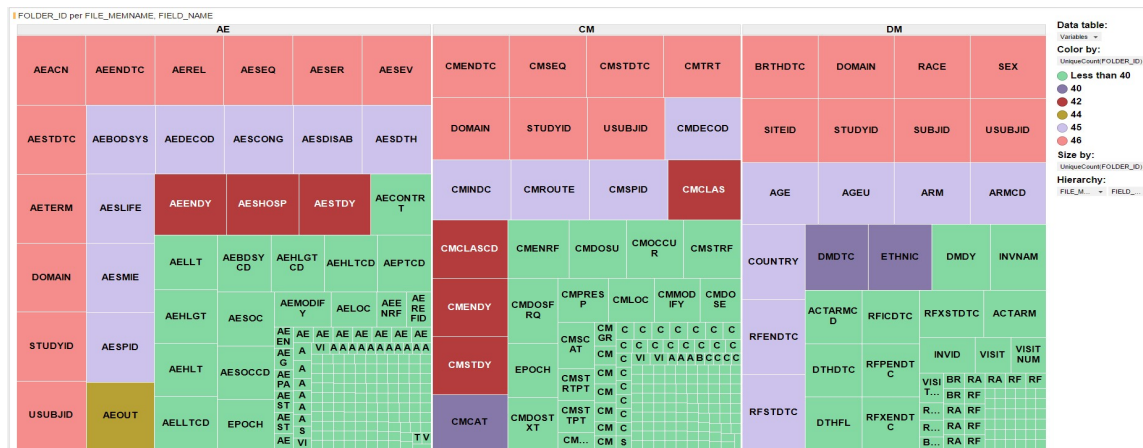


Figure 6. Variable selection capability. Variables colored in light-red are present in all 46 selected studies. Select variables (for analysis) and query SAS data to populate analysis results in Figure 7.

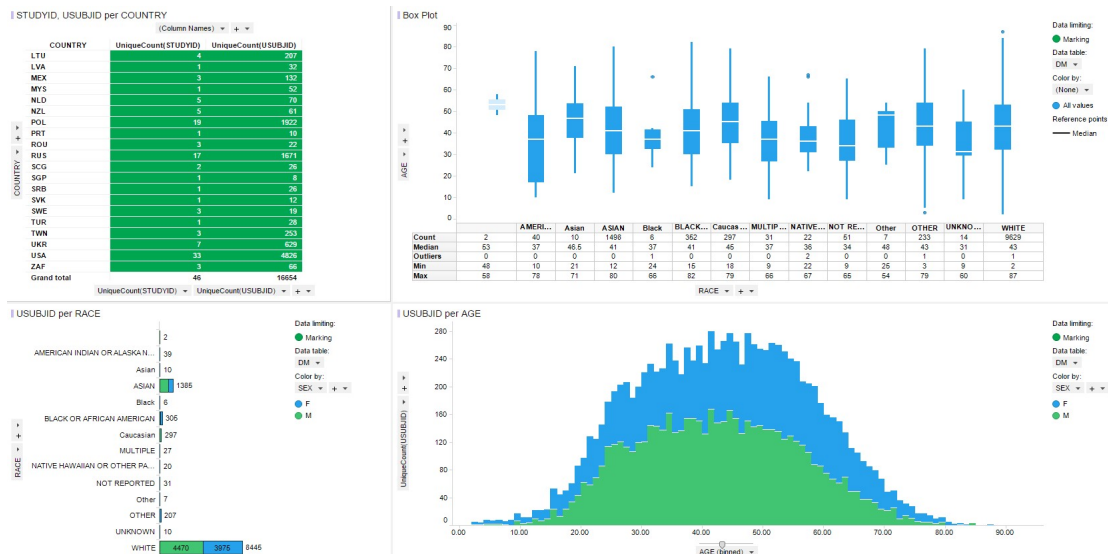


Figure 7. Analysis results of SDTM (DM) data profiling for 46 studies from 3 compounds.

CHALLENGES

Data Standards Conformance: Figure 8 shows 3 types of data (Analysis, Tabulation, Raw) in PDW. Analysis data type is shown in blue and it is seen that more than 50% of the analysis data conform to ADaM model, and the rest of the analysis data conform to the in-house defined Analysis Data Model (ADM), Harmonized Analysis Data Set (HADS), or to legacy structures. The Tabulation data type (in red) contains data conforming to SDTM data model, and the Raw data type (in green) contains CRF/collected data conforming to the in-house defined Common Raw Data Model (CRDM), or to legacy structures.

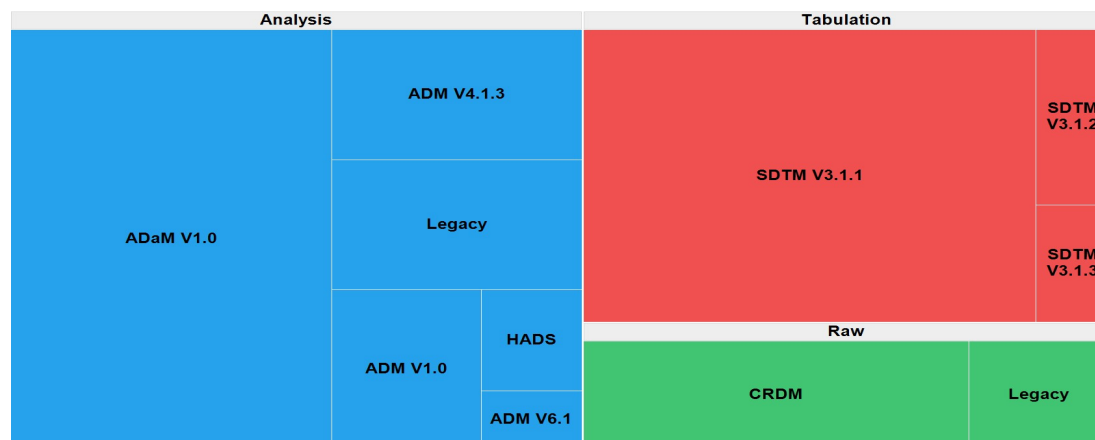


Figure 8. Distribution of data standards across analysis, tabulation and raw data types in PDW.

Data Mapping/Transformation Efforts: Due to the fact that PDW data is *loaded as is* based on the original SAS datasets created for individual studies without cross-study data standardization/harmonization considerations, most of the PDW use cases support required data mapping and transformation. The following shows the challenges from a PDW use case support which creates pooled ADSL and ADEFF from 10 studies with the same indication:

1. Same information created in different places: ADSL Baseline XYZ Dose (obtained from PDW metadata). Table 1 demonstrates this case from a PDW use case where ADSL and ADEFF datasets were pooled from 10 studies with the same indication.

Study ID	Data Source	Dataset	Variable	XYZ Dose Range
S0001	SDTM	CM	CMDOSTXT	(2,25)
S0002	SDTM	CM	CMDOSTXT	(7.5,25)
S0003	ANALYSIS	MEDRVIEW	MRXYZDS	(0,...,30)
S0004	ANALYSIS	EXPHX	EHSCRDOS	(15,25)
S0005	ANALYSIS	SUBJSF	XYZDOSBL	(0.5,30)
S0006	ANALYSIS	MEDRVIEW	MRXYZDS	(0,25)
	ANALYSIS	SUBJEF	XYZBL	(15,25)
	ANALYSIS	EXPHX	EHSCRDOS	(15,25)
S0007	ANALYSIS	ADCM	CMDOSTWK	(7.5,20)
S0008	ANALYSIS	ADSL	XYZBL	(12,5,15,17,...,25)
S0009	ANALYSIS	WNLOBS	XYZD	(10,25)
	ANALYSIS	EXPHX	EHSCRDOS	(10,25)
S0010	ANALYSIS	ADSL	BLXYZDS	(10,25)

Table 1. ADSL Baseline XYZ Dose created in different datasets and under different variables.

2. Same variable name with different derivation applied: The ADaM Change from Baseline variable CHG is defined as CHG=AVAL-BASE in some studies and BASE-AVAL in others. The integration of such analysis data is a challenge. Best practices to ensure integrity of data integration need to be defined and reproducibility of original datasets should be a litmus test.
3. Variables not created in analysis dataset: The focus on a single clinical study may lead to the creation of a limited set of variables. As a result, many common variables that are needed for data integration and for the larger cross study analyses may be missing, leading to data integration challenges.



4. Lacking metadata or variable flags for duplicate records: Without straightforward/direct data selection algorithm of duplicate observations/records for analysis is commonly observed in this PDW use case. For example, duplicate measurements for the same analysis visits but without appropriate variables such as DTYPE and/or analysis flag variables such as ANLxxFL to facilitate the unique record selection.

Extra efforts for verification: To ensure the accuracy of the final analysis results, verification tools for both analysis data and outputs are created due to the data manipulation above. When many studies are involved, report comparisons are cumbersome and time consuming.

CONCLUSION

The new capabilities to search, query and perform analysis across hundreds of studies on the PDW platform is a major achievement which substitutes the struggles of traditional cross-team manual and fragmented efforts. The platform enables the automation of data integration and analysis by providing the user with a friendlier experience while supporting time has shortened tremendously.

Lack of conformance to data standards is the biggest barrier to meaningful data integration for analysis. Improving on this process is one of the major tasks moving forward. Technological advances are accelerating and provides us with options to address these challenges. Artificial intelligence and machine learning techniques are making their way into clinical trials end to end processes. In collaborating with vendors, we are evaluating new waves of the data standardization tools. In addition, we are planning to expand the scope of PDW data from completed studies to ongoing studies as data becomes available. The experiences from PDW as well as designs and technologies will be employed in cross-study data standards monitoring to ensure data standards conformance to SDTM/ADaM at all metadata levels with the intention to perform real-time analysis in the end-to-end drug development cycle.

As PDW data content increases, we also continue to improve our technological solutions to address performance issues.

REFERENCES

Fielding, R, Representational State Transfer (REST)
https://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm
Reuben, A, 2004, Hy's Law, Hepatology, 39(2):574-8.

ACKNOWLEDGMENTS

Quantitative Sciences (QS), Janssen Pharmaceutical Company of Johnson & Johnson
Integrated Data Analytics & Reporting (IDAR), Janssen Pharmaceutical Company of Johnson & Johnson
Global Technology Services, Johnson & Johnson

CONTACT INFORMATION

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

Sandy Lei / Simson Alex
Johnson & Johnson
1165 Trenton-Harbourton Road
Titusville, NJ08560
Work Phone: +1 732-698-8524 / +1 484 886-6198
Email: slei@its.jnj.com / salex2@its.jnj.com