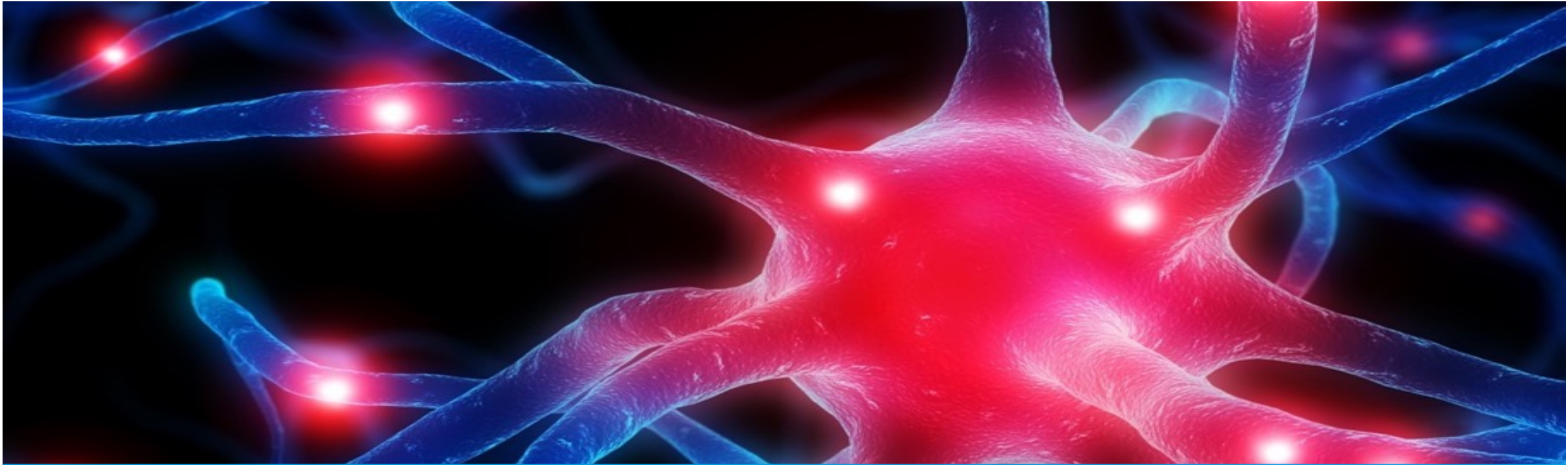




Programming Infrastructure to automate Process flow for Central Statistical Surveillance

Presentation No: AR03

Archana Nayak
Salvatore Saporito
Kris Lauwers
Stefan Van Yper
Mateusz Hinzmann
Bhargava Reddy

Active neuron

Agenda



**Introduction
to CSS**



**Current state
of CSS**



**Process flow
of CSS**



**Programming
infrastructure**



Future steps

Central Statistical Surveillance



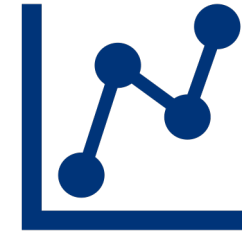
What?

Launched in 2019

Internally developed capability

Part of Analytical Risk based Monitoring (ARBM) offering

Important layer of clinical trial data oversight



How?

Looks holistically at clinical trial data

Employs statistical models to detect naturally occurring data patterns and relationships

Identifies outlying sites breaking anticipated patterns

CSS overview

- CSS augments Analytical Risk based Monitoring
- Additional layer of protection for subject safety and data quality
- CSS high level process map is as highlighted in the figure



Figure 1. High Level Central Statistical Surveillance Analysis Process Map.

CSS is a collaborative effort

Risk Management – Central Monitoring



Marion Wolfs;
Head, RMCM



Susan Meier;
Director,
RMCM



Mateusz
Hinzmann;
Manager,
RMCM



Marcelina
Markiewicz



Jan
Gasiorowski



Anna
Radziszewska



Weronika
Stepczyk

CSS Analysis Team

Advanced Insights



Archana
Nayak



Kris Lauwers



Salvatore
Saporito



Stefan Van
Yper

Statistics and Programming Team



Eileen Doherty
Head, Clinical
Operations
Solutions



Bhargava
Reddy; Director,
Advanced
Insights

CSS Capabilities

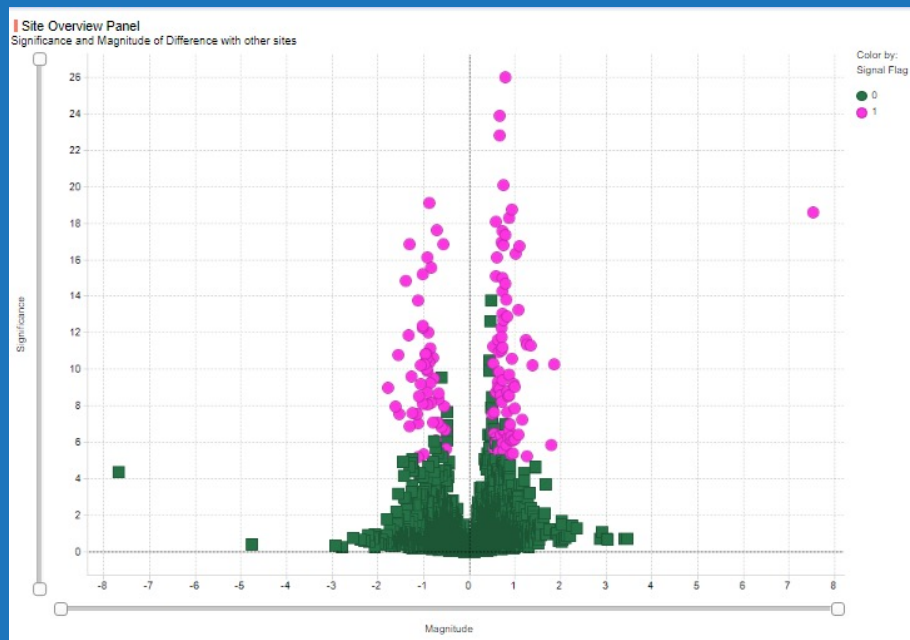
Following models are implemented within CSS analysis:

- **Average Test Value Analysis**
- **Between Subject Analysis**
- **Within Subject Analysis**
- **Between Collection Timepoint Analysis**
- **Within Collection Timepoint Analysis**
- **Categorical Data Analysis**



Average Test Value Analysis

Identification of sites that differ from the normal study data profile:

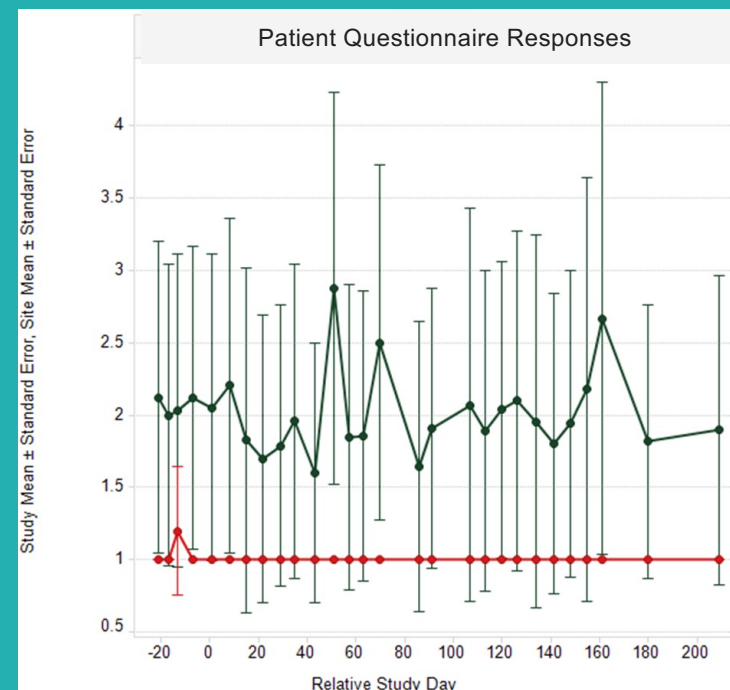


LEGEND:

- **Dot Color:** darker pink color indicates degree to which the site's data differs from all sites in the study

Analysis includes EDC, questionnaire, and lab data for the qualified sites.

Results at site in question are more alike compared to the study norm:



LEGEND:

- **Green Line and Error Bars:** patient assessment results across all sites in study
- **Red Line and Error Bars:** site in question showing results differing from study norm and with almost no variation/error bars

Between and within Subject Analysis

Between Subject Analysis shows how much subjects differ from each other. It gives an estimate of the variability between subjects at site level and compares that with all other sites

The analysis demonstrates whether subjects at a site are more alike (low variance) or more different (high variance) than expected

Differences in **Within Subject Variance** tell us about how subject test results changes **over time**

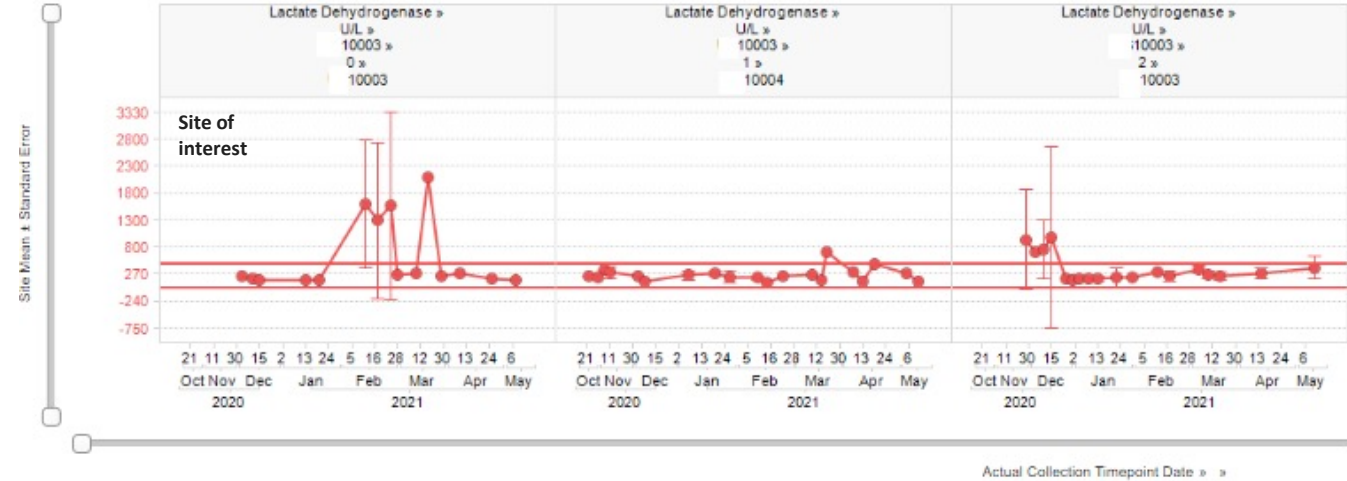
Site exhibiting dynamic change in parameter over time (subject have extreme value of parameter, much above biological range)



Between and within Collection Analysis

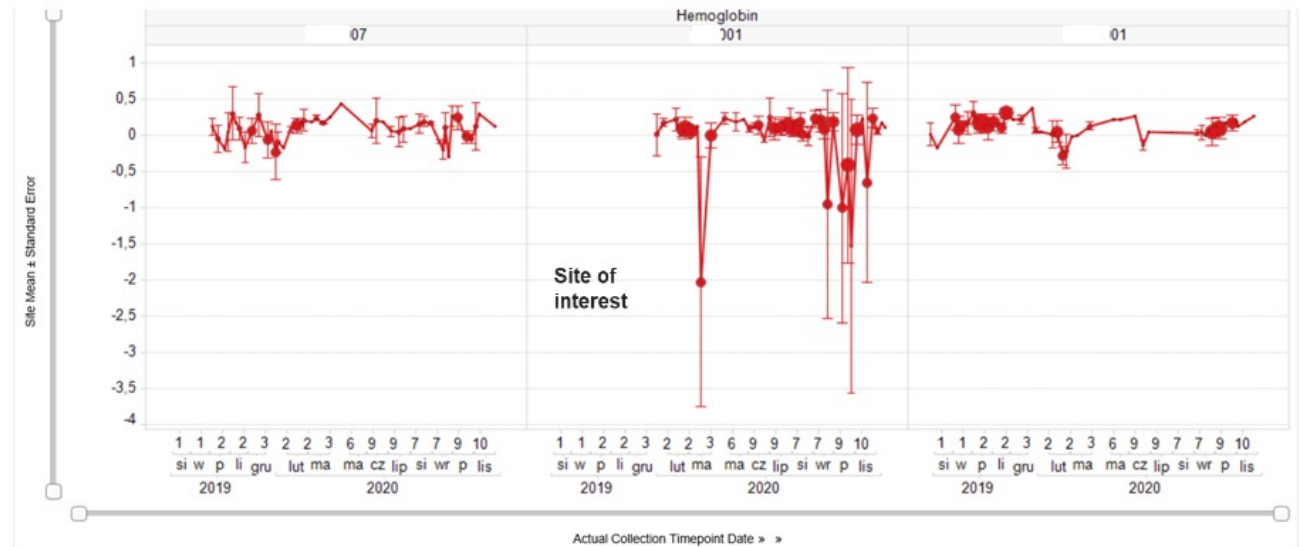
Between collection timepoint analysis shows the change over time in test results at a site when the test results are grouped based upon the **date of collection**

Example of site of interest of potential finding with high variability, compared with sites of similar number of subjects enrolled



Within collection timepoint analysis shows the degree of the difference of a test result grouped and assessed by collection date at a particular site

Example Site of interest exhibits high variability in the value of parameter at a **particular timepoint**

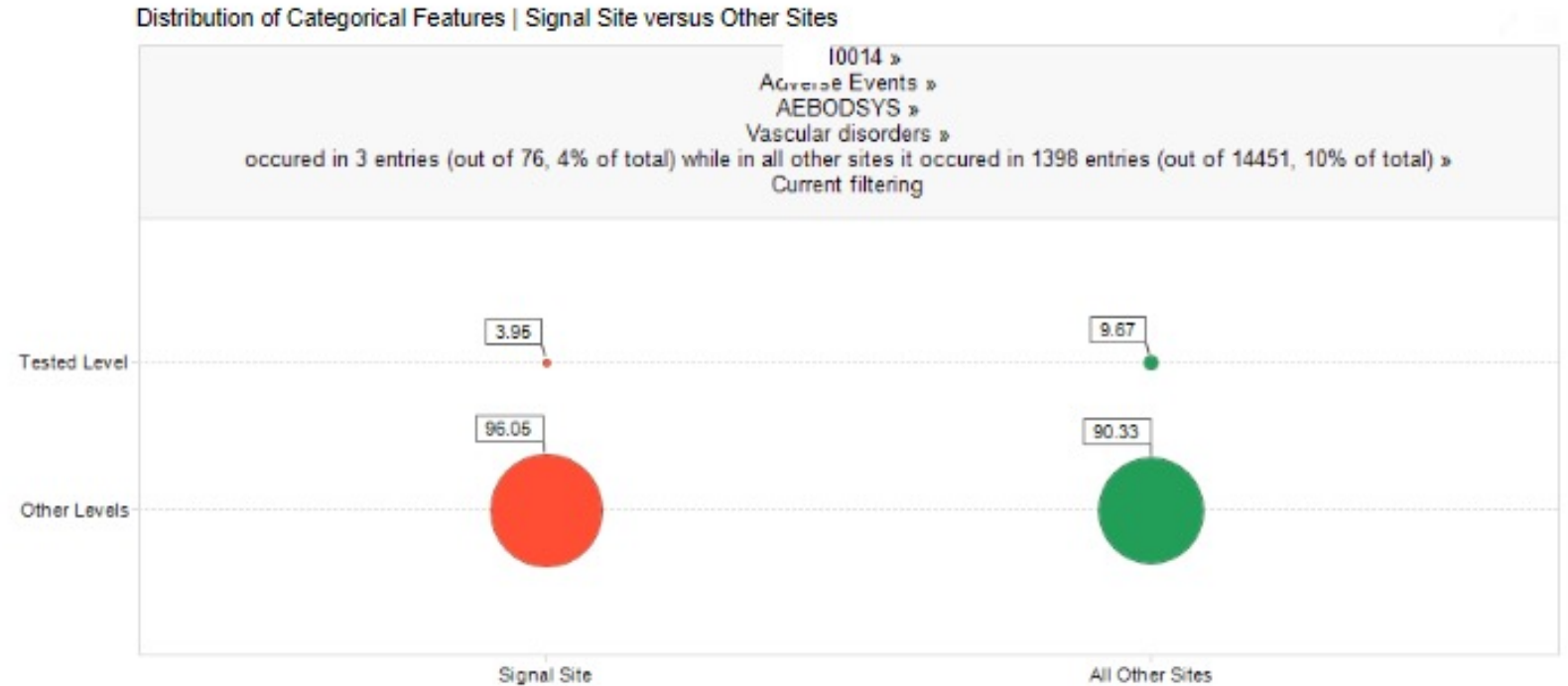


Categorical analysis

Categorical variables: nonnumeric quantities with nominal values or ordinal values

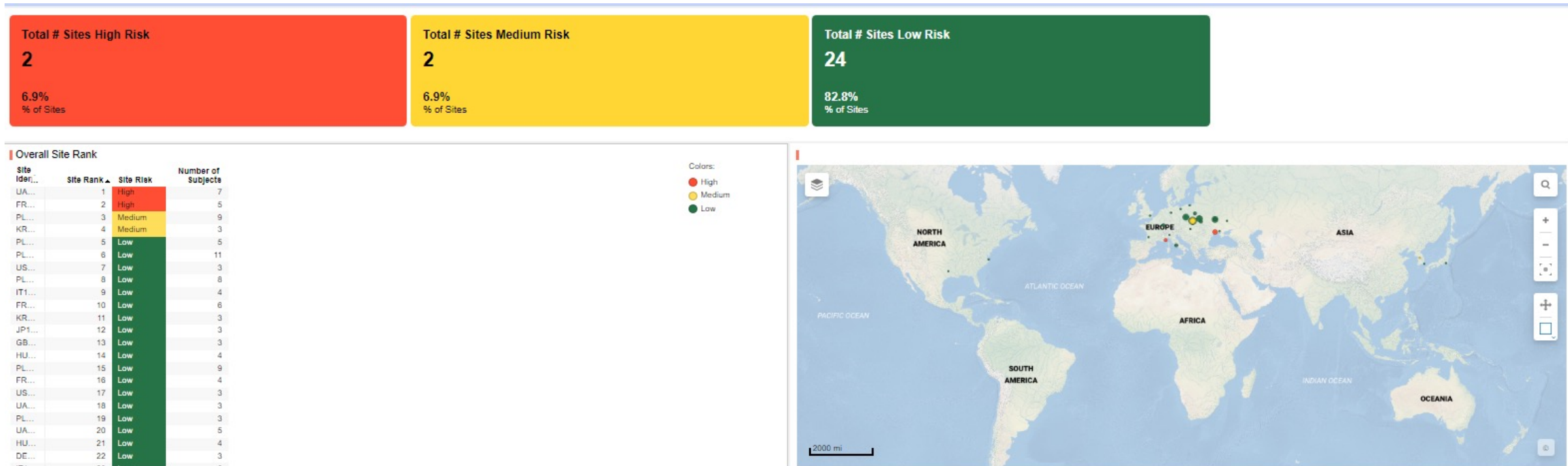
Categorical variables are characterized by limited, and usually fixed, number of possible values (levels)

Deviations are identified by testing the homogeneity of probability distributions; between sites at individual levels



Next Generation features

- ✓ Simplified dashboard to allow efficient signal detection and reporting
- ✓ Combine significance and magnitude of a test result to define a relevant signal
- ✓ Overall ranking mechanism to highlight sites of interest
- ✓ Pre-identified signals allowing for quasi single click report creation
- ✓ Addition of report page gathering all the identified signals



Operational analytics and alliances

 **101+** data processing cycles for  **44+** Janssen studies across  **7** therapeutic areas



Oncology



Immunology



IDV
Infectious diseases
and vaccines



Neuroscience



CVM
Cardiovascular medicine



Medical
Affairs



Actelion

Scope and outcomes

1000+ outliers identified amongst multiple sites



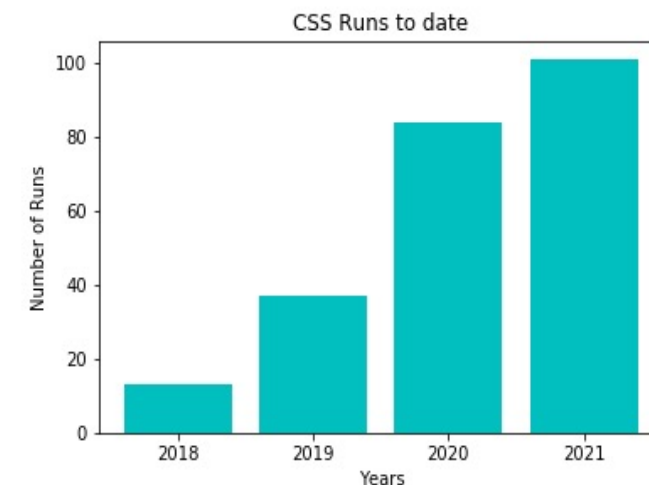
160+ potential findings identified as sites of interest for further investigation



40+ study teams informed with near real time data insights

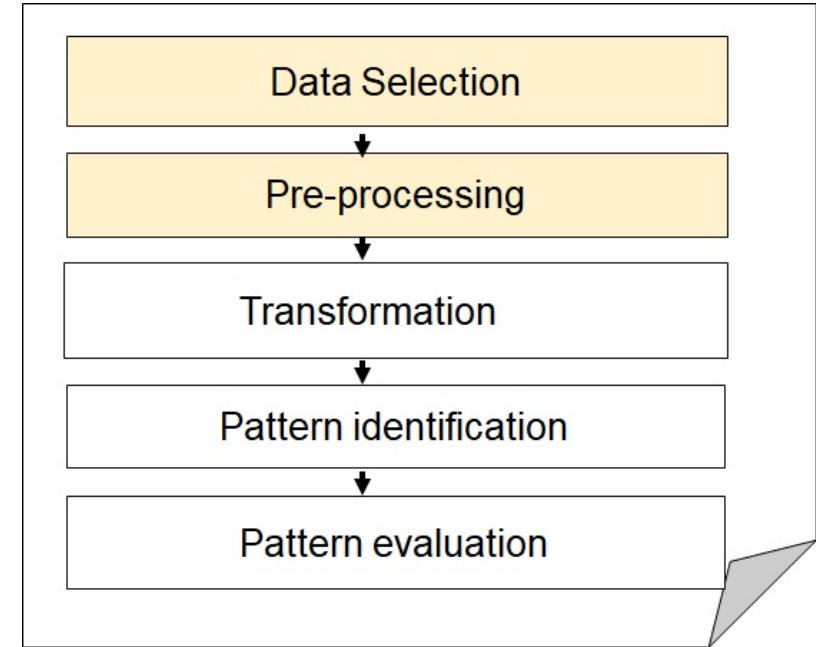


60+ potential findings translated into action



Goals

- Increase **scope** of analysis by automation of data ingestion
- Increase **accuracy** with lesser manual interventions
- Operational **efficiency** gain by decreasing data processing turn-over time



Overview steps of data analysis

Challenges of data pre-processing steps

- **Different therapeutic areas;** study can have variety of datasets based on therapeutic areas
- **Heterogenous data formats;** CDSIC standards adaptation between studies
- **Variability;** study specific annotations for test codes, relative time stamp calculations, datasets selection and formats
- **Programming infrastructure;** machine hardware and memory for tasks

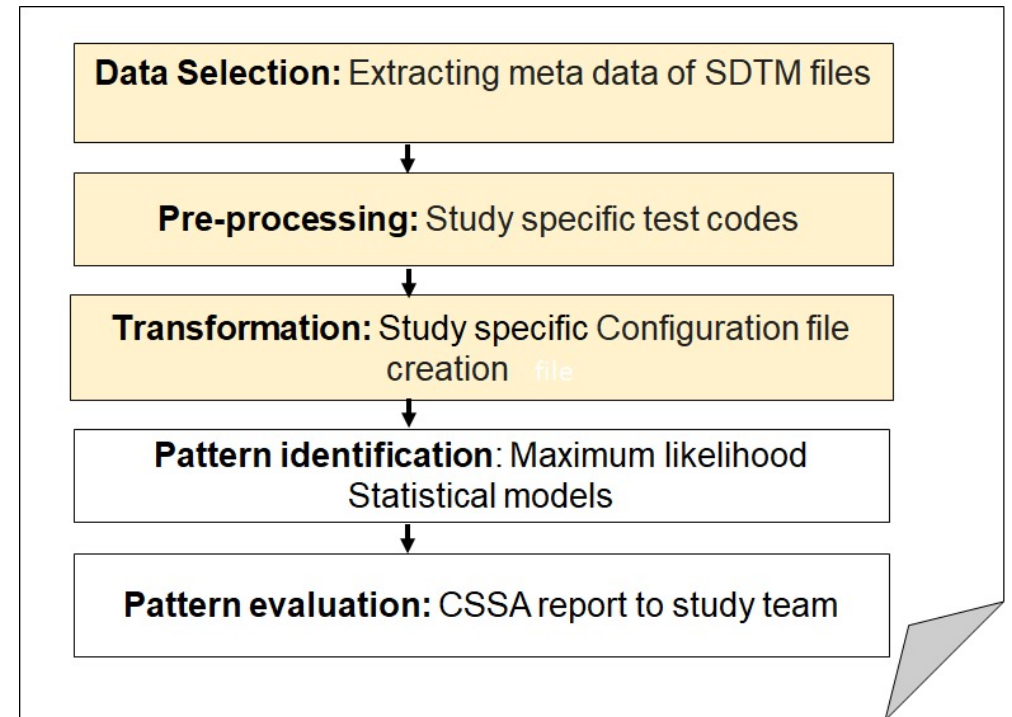


Solutions built (Process flow for CSS)

Method:

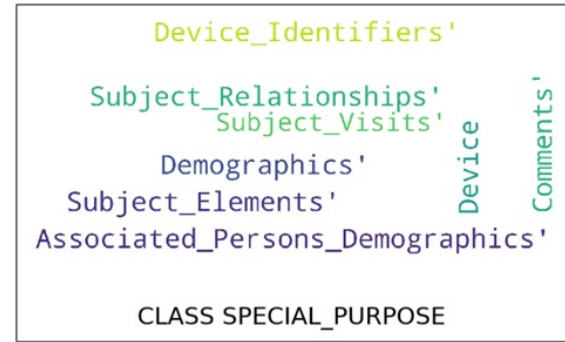
Automation using python and R programming by:

- Eligibility analysis of first run set up
- Extracting meta data of SDTM files
- Study specific test codes
- Configuration file creation
- Pre-run review with study team
- Integration with cloud platform
- Git version control for track changes

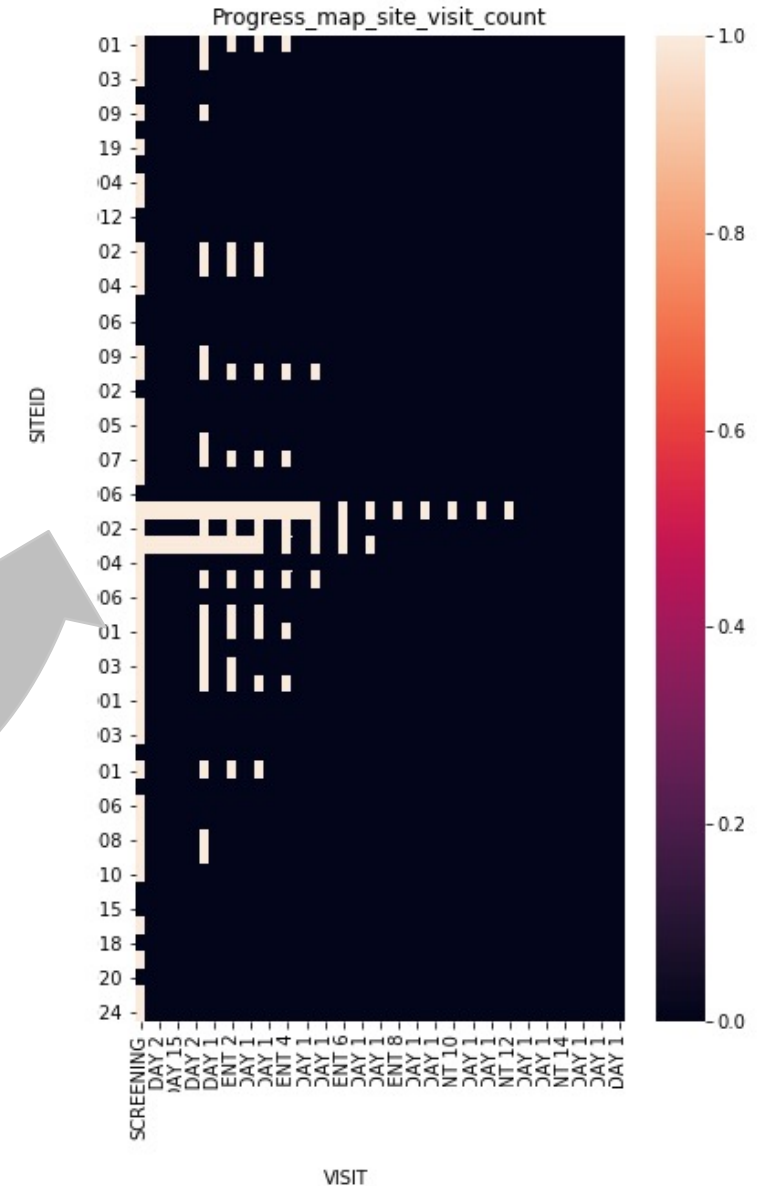


Eligibility analysis

- CSS mixed linear model requires sufficient sample size for convergence of the maximum likelihood optimization
- Eligibility criteria for CSS analysis; study data with at least 20 sites and at least 3 subjects per site and with at least 3 visits per subject



In the figure, Heat map shows an estimation of qualifying sites (vertical axis) and number of visits (horizontal axis) set to scale limits between 0 (low) and 1 (high) as analyzed from SV and DM data domain datasets.



Extraction of meta data from SDTM

- Automated script to extract the different datasets present in study folder of interest
- Evaluating the test codes, number of missing values, present and unique values
- Filtering domains such that those without test code will be excluded, test differentiators are built-in, etc.

<u>Variable names</u>	<u>Missing values</u>	<u>Present values</u>	<u>Unique values</u>
STUDYID	0	936	1
RDOMAIN	0	936	1
USUBJID	0	936	486
IDVAR	0	936	1
IDVARVAL	0	936	16
QNAM	0	936	1
QLABEL	0	936	1
QVAL	0	936	2
QORIG	0	936	1
QEVAL	0	936	1

<u>Domain variables</u>	<u>Missing values</u>	<u>Present values</u>	<u>Unique values</u>
raw file	/mnt/STUDY_ID/data/lb.xpt	/mnt/STUDY_ID/data/lb.xpt	/mnt/STUDY_ID/data/lb.xpt
STUDYID	0	54593	1
LBSPID	0	54593	
LBTESTCD	0	54593	95
LBTEST	0	54593	95
LBCAT	0	54593	4
LBSTRESC	0	54593	2660
LBSTRESN	7604	47535	2539
LBSTRESU	0	54593	10
LBSTNRLO	6603	48536	44
LBSTNRHI	4949	50190	58
LBSTNRC	0	54593	5

Study specific configuration file

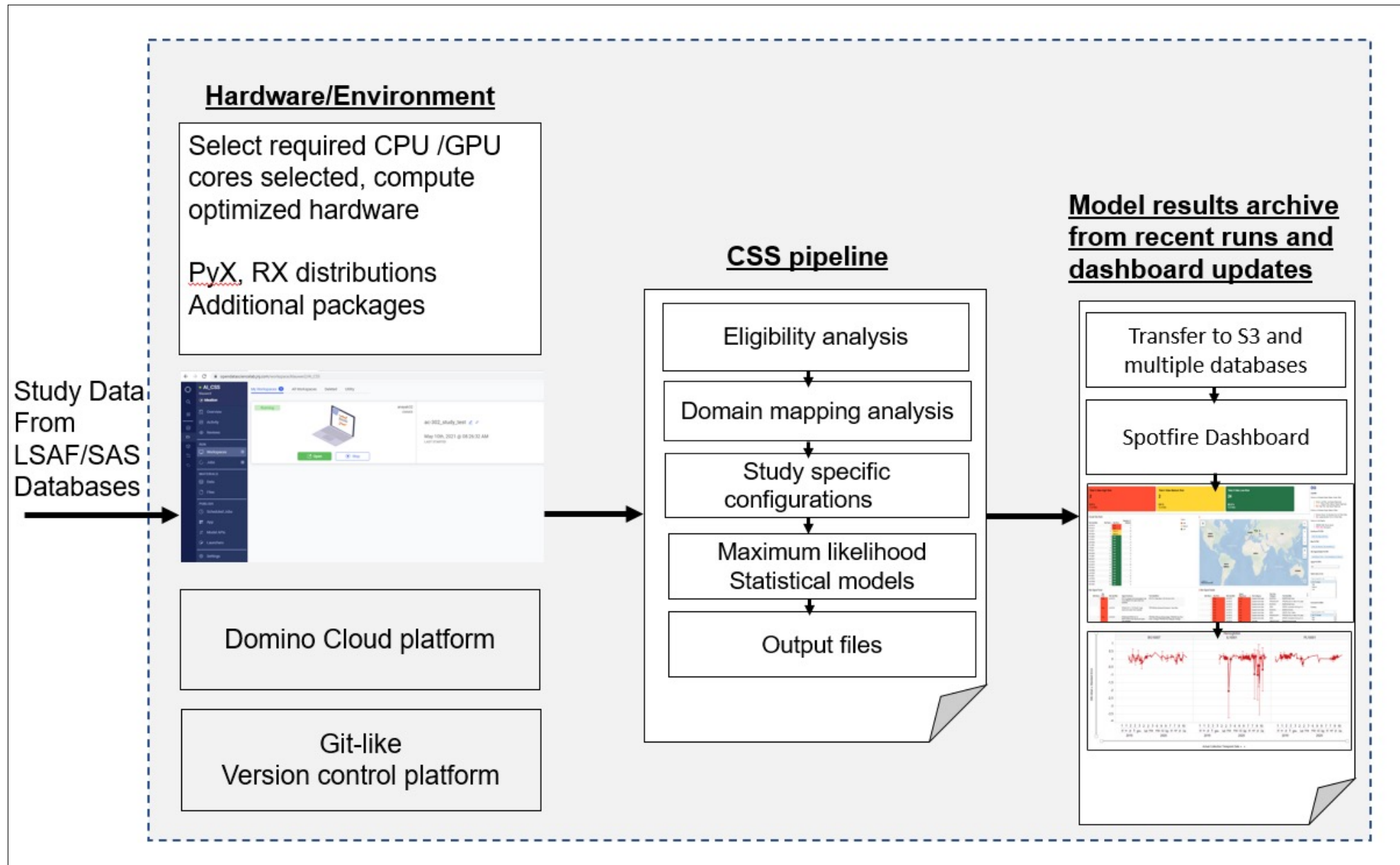
- Numerical and categorical test values separated into sets
- Visit schedule extracted and test with multiple visits included
- Duplicated test codes across domains are appended to create list of unique tests with appropriate unit of measurements
- Mapping SDTM datasets to EDC, questionnaire, and lab domains for the qualified sites

TEST	VISIT	DACAT	QSCAT	TRCAT	LBCAT
DISPSTAT	50	3			
ECOG101	83		1		
DIAMETER	7			1	
BILDIR	99				1
BILI	105				2
GGT	104				1
GLUC	105				2
HCAB	7				1
HCG	4				2
RBC	104				2
HIVAB	6				1

Test Name	VISIT	MI counts	SR counts	LB counts	Criteria
ULCER	8	5			INCLUDED
INDURD IA	1		5		EXCLUDED
INTP	2		0		INCLUDED
BASO	36			29	INCLUDED
BASOIM	1			0	EXCLUDED
BASOIM LE	1			0	EXCLUDED
BASOLE	36			31	INCLUDED

LBTEST CD	LBCAT	LBSPEC	LBSTRESU
RBC	URINALYSIS	URINE	/HPF
RBC	HEMATOLOGY	BLOOD	10^12/L
BILI	CHEMISTRY	BLOOD	umol/L
BILI	URINALYSIS	URINE	-

Programming infrastructure



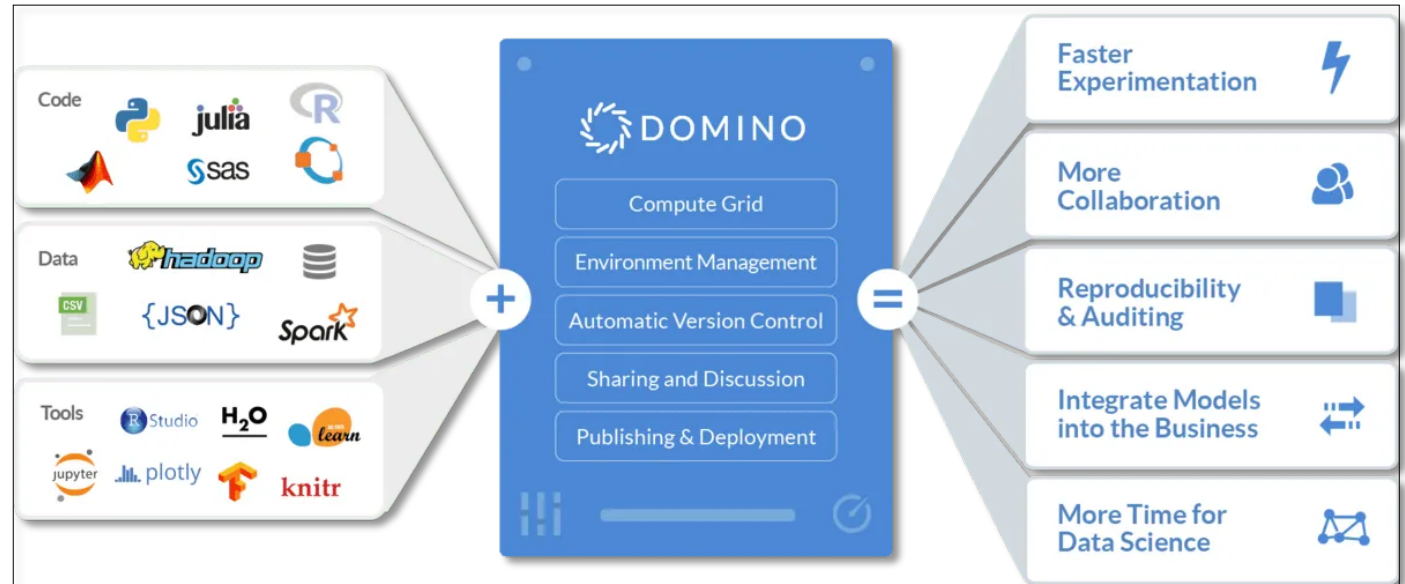
CSS operational runs carried out on Janssen Open Data Science Lab platform (Domino)

CSS runs on Domino platform

Efficiency: Faster Turnaround time from bug/feature identification to production

Collaboration: Allow for parallel development by multiple collaborators

Reliability: Track changes with commit history



Summary

Scope

- ✓ Automation of Study specific configuration files to streamline the process
- ✓ Lesser manual interventions

Accuracy

- ✓ Process of SDTM datasets selections in Class findings is fastened due to auto-evaluations
- ✓ Communication a priori with the study team has allowed greater transparency
- ✓ Study specific configuration file serves as reference file during dashboard results analysis

Efficiency

- ✓ Reduce review time
- ✓ Users report an efficiency gain (speed of analysis completion) of up to 30%
- ✓ Next generation features of dashboard has allowed faster report generation for study teams

Acknowledgments

Many thanks to PHUSE EU CONNECT 2021 for this opportunity

Many thanks to all Janssen and ICON plc

Especially all team members of Advanced insights, Data science, RMCM (ARBM)teams

Special thanks to Central Data science platform team (Domino) and

R&D Advance Computing- Data Analytics & Engineering teams

FLEX PRAHS/ICON PLC



Stronger together. The new ICON.



PRA Health Sciences has been acquired by ICON plc, creating the world's most advanced healthcare intelligence and clinical research organization.