



Central Statistical False Positives – Predicting True Signals with Machine Learning

Dorothea L. Ugi, Manager Statistical Programming

The Janssen Pharmaceutical Companies of Johnson & Johnson
PHUSE US Connect 2021

June 14th – 18th, 2021

Cliff Enright, *Star Cradle*

Artwork from The Creative Center at
University Settlement.

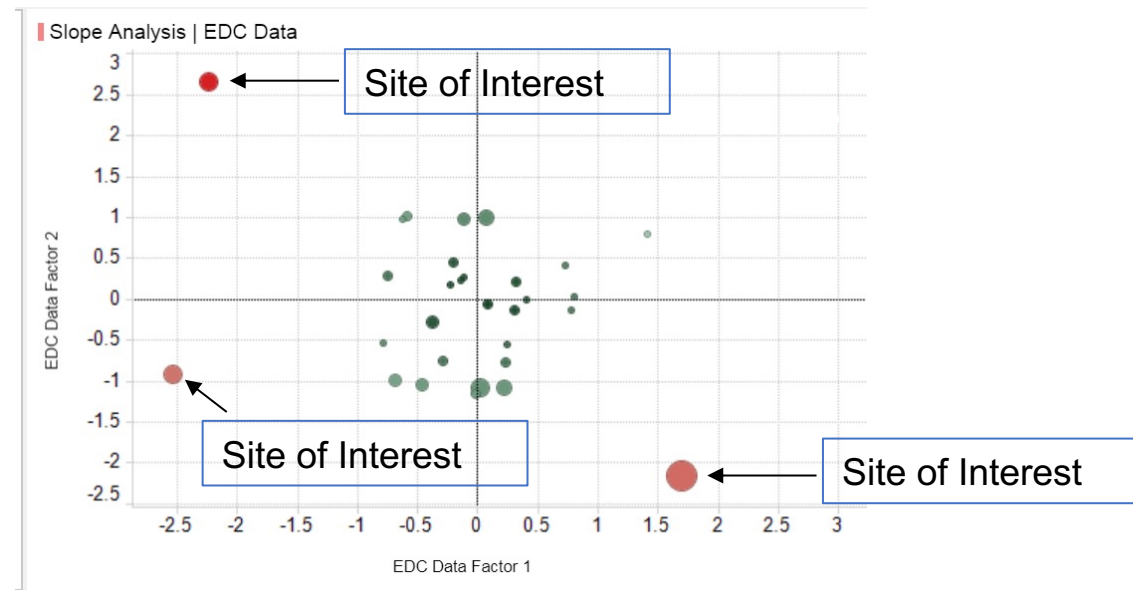
Agenda

- **Central Statistical Surveillance (CSS)**
 - What is it?
 - Interesting Findings
 - Benefits & Success Story
- **Improvement Opportunity**
 - Objectives & Definitions
 - Summary of Data & Variables Used
 - Training Set
 - Results
 - Applying Prediction Probabilities

What is Central Statistical Surveillance (CSS)?

The methodology of Central Statistical Surveillance is to determine the naturally occurring relationships within a trial's data and identify the differences. Naturally occurring relationships are determined for all data within a trial. These “Naturally occurring relationships” on the study level are used as the comparator for the site level relationships so that sites that differ can be identified.

DATA = Questionnaires, Lab & EDC (must have visit date and numerical value)



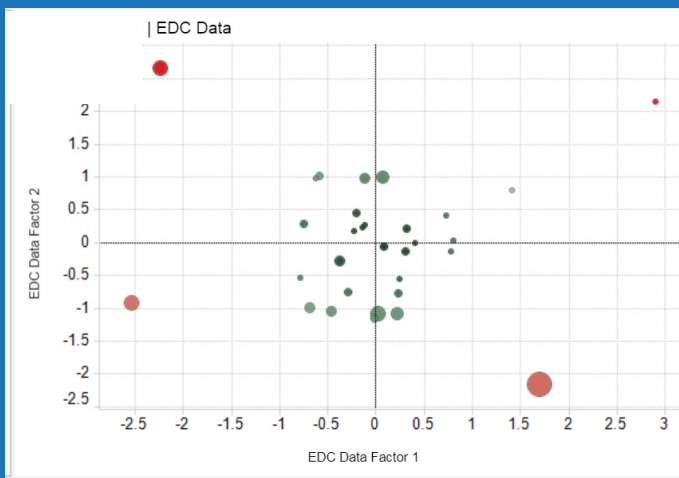
Development



- **Central Statistical Surveillance** (CSS) is an innovative, sophisticated statistical capability that provides another layer of oversight.
- The development started as an augmentation of the risk management-central monitoring tools.

Visualizing Sites against the Study Norm

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

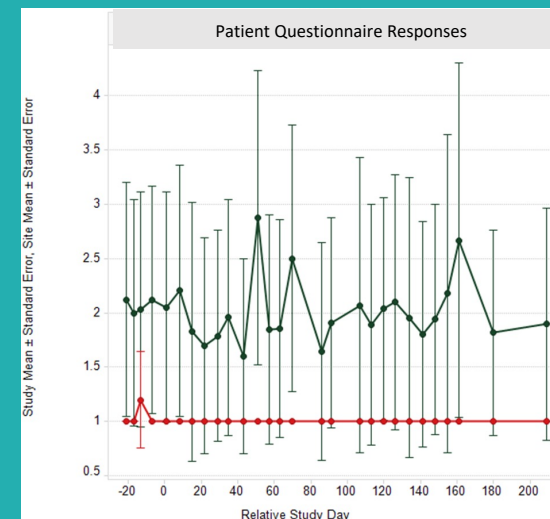


LEGEND:

- **Dot size:** number of subjects at site
- **Color:** darker red color indicates degree to which the site's data differs from all sites in the study

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

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

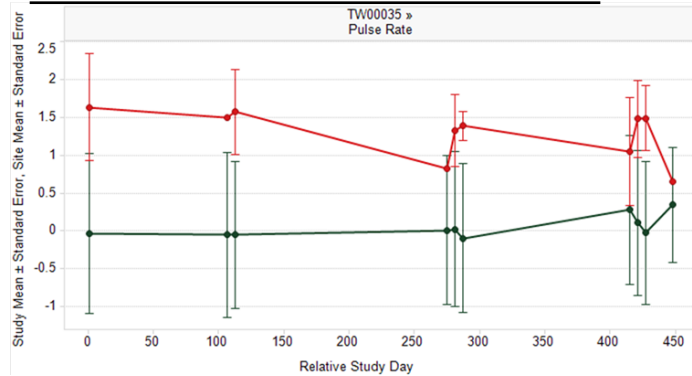


LEGEND:

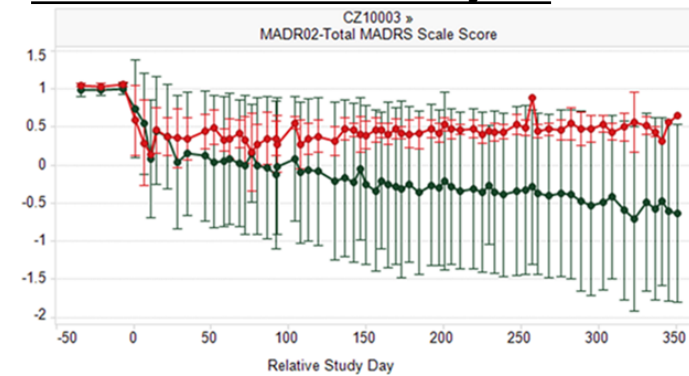
- **Green Line and Error Bars:** patient assessment results across all sites in study
- **Red Line and Error Bars:** site in question showing results differing from study norm and with almost no variation/error bars

Analysis Methodology

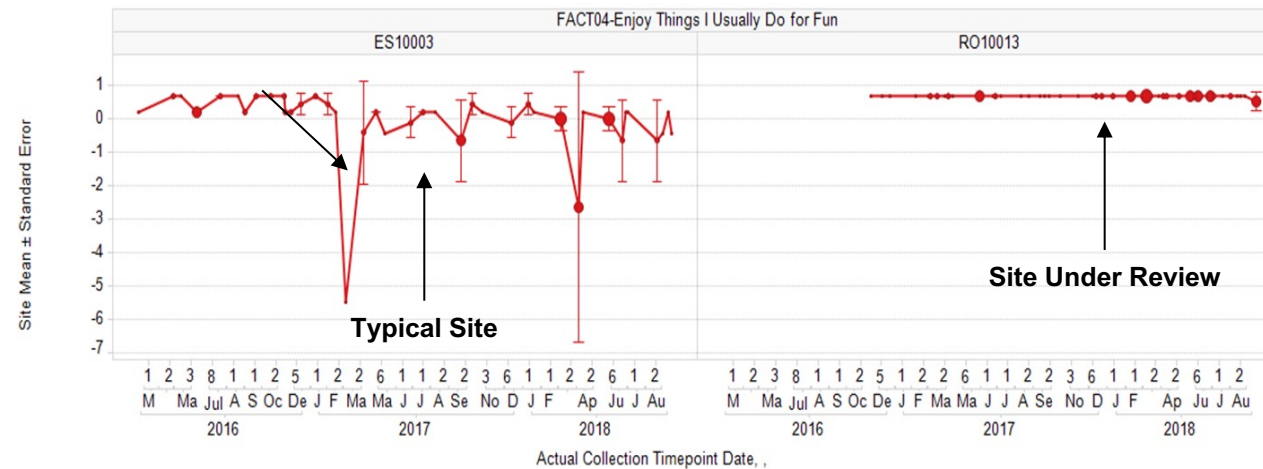
Differences between test results



Differences between subjects

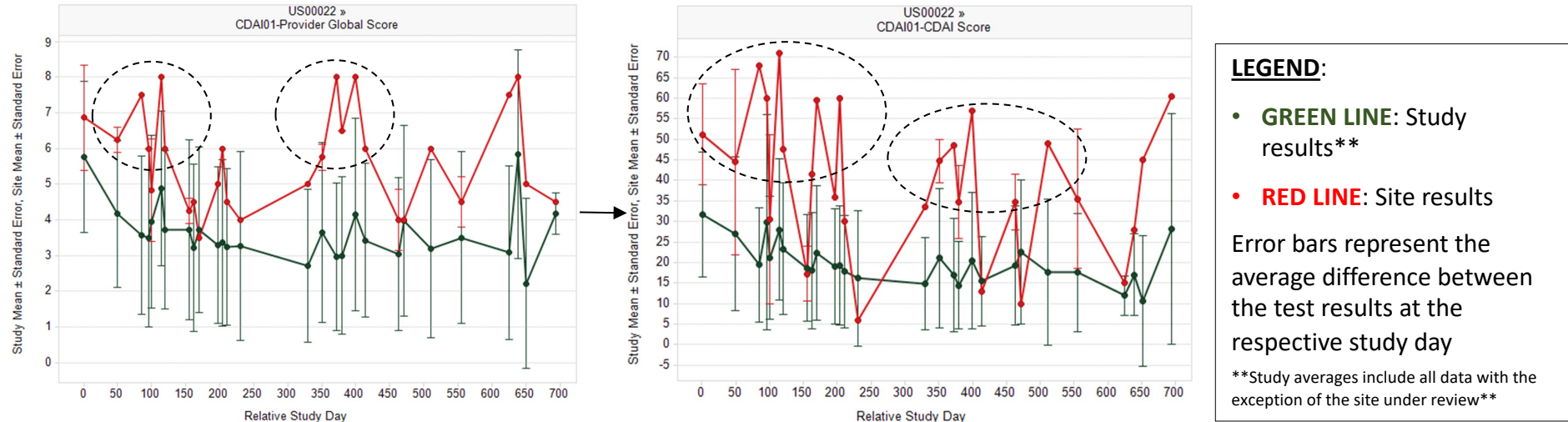


Differences in test results based upon their collection date



Observation Examples

Example Observation #1: CDAI results are consistently greater than those at a typical site



- CDAI score results are significantly higher than the study average.

Recommendations to Study Team:

- Review site level CDAI results
- Assess site process for administration/collection/calculation of the CDAI

Finding after Study Team Review:

- Site calculated CDAI score prior to entry into EDC, where the CDAI score is imputed within EDC.

Example Observation #2: Lab results are more alike than the study average

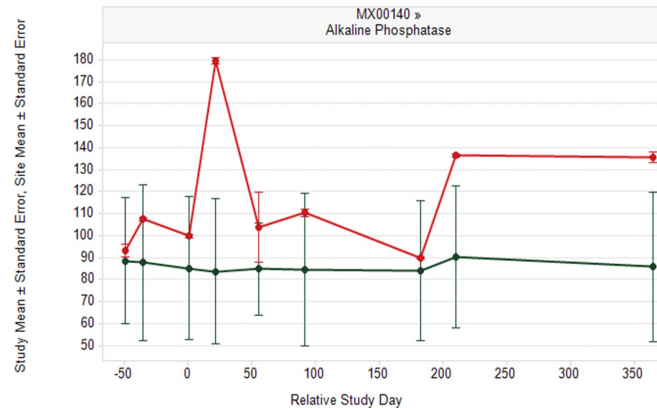
- Test results are very similar on both study days and actual collection dates
- Subjects appear to have all been brought to clinic on the same dates

Recommendations to Study Team:

- Review site level laboratory results to determine if there are any clinical issues
- Review source notes to ensure batching of subjects/samples was not an issue

Finding after Study Team Review:

- Site utilized the same blood samples for all subjects at all visits

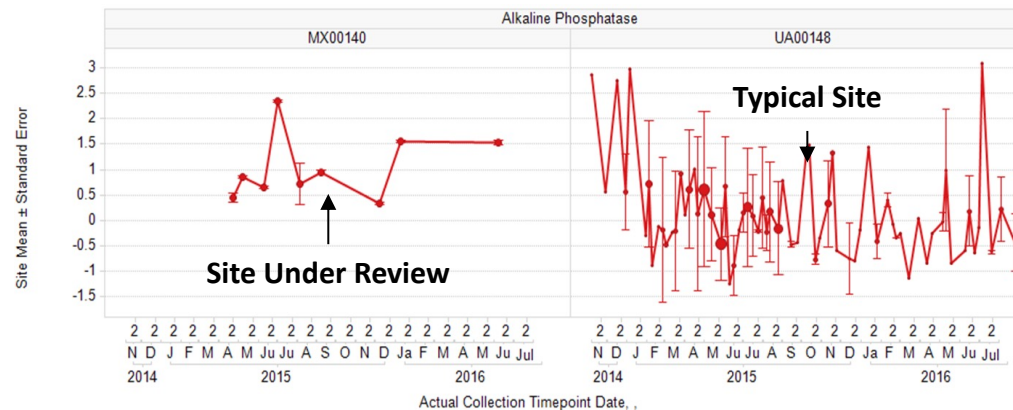


LEGEND:

- **GREEN LINE:** Study results**
- **RED LINE:** Site results

Error bars represent the average difference between the test results at the respective study day

Study averages include all data with the exception of the site under review



LEGEND:

Site under review and comparator sites visualized for comparison.

- **Dot size:** number of samples collected at the timepoint
- **Error Bars:** difference between test results at a given point

Benefits



Central Statistical Surveillance augments Risk Based Monitoring looking at data holistically using a multivariate approach.



Ability to identify difficult to detect data anomalies and systemic issues.



Additional layer of protection for **subject safety** and **data quality**.



Early detection of critical data issues that could jeopardize overall study data.



FDA employing similar methodology for site inspection identification.

Success Story:

Health Authority Site Inspection Identification Alignment



- Scope: Phase 3 study.
- Result: Central Statistical Surveillance activities identified 7 of the 10 sites identified for Inspection by the FDA.
- Conclusion: Unintentional data errors were often a predictor of additional site quality issues.

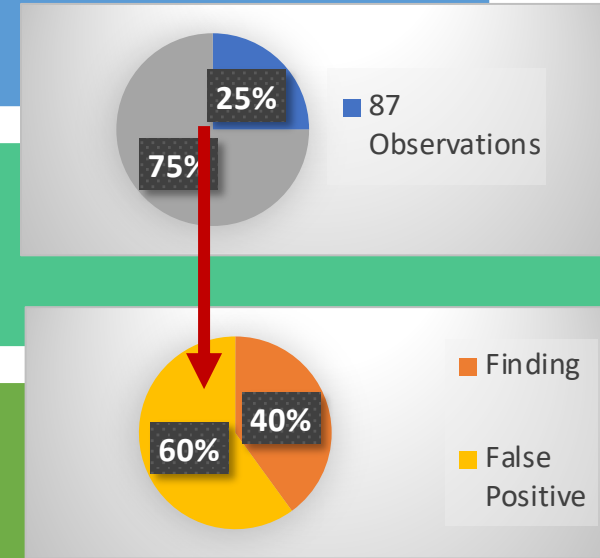
Quality Risk Management Metric Insight

- Scope of 22 runs
- Across 18 studies
- In 4 therapeutic areas

350 Site Outliers

87 Observations

34 Findings



Conclusion: 60% False Positives

Objectives: Reducing False Positives

Business Problem

- Current methodology incorrectly flags 60% of the sites in a study as high-risk site.
- Investigating signals is a very manual and time-consuming process for the analysts.

Solution

- Automated identification of sites using Machine Learning methodology.
- Predict True Signals (TSs) from all the signals identified which would reduce the amount of time analysts requires to fully review a study.

How do we solve it?

- We predicted if a site has any True Signals(TSs) using classification methods.
- This will allow for targeted monitoring.
- Reduces the number of signals an analyst needs to investigate.

Definitions

- **General Linear Model (GLM)** is a regularized method that tries to balance the model performance and the complexity of the model. The two methods, ridge regression and lasso regression, are utilized to optimize the specific loss function using all the available data in the learning sample.
- **Gradient Boosting Model (GBM)** is a non-parametric tree-based ensemble model that has been developed to solve classification- and regression-type problems.

Summary of Data Included

- Small sample size
 - 1418 Sites
 - 22 Studies
 - 2 Phase II & 20 Phase III
- 4 TAs
 - 7 Oncology (469 Sites)
 - 7 Neuroscience (474 Sites)
 - 5 Immunology (389 Sites)
 - 3 ID&V (86 Sites)

TA (Phase)	STUDYID	Compound	Indication	For TA (#)	# of Sites	# of Sites	anySignal		# of Sites	anyTS		# of Sites	Only FS	
							For TA (%)	#		For TA (%)	#		For TA (%)	#
ID&V	63623872FLZ3001	63623872	Influenza	86	21	10	41(47.7%)		2	10(11.6%)		8	31(36.0%)	
ID&V	63623872FLZ3002	63623872	Influenza		31	17			6			11		
ID&V (2)	56136379HPB2001	56136379	Hepatitis B		34	14			2			12		
Immunology	CNT01275SLE3001	54160353	Systemic Lupus Erythematosus	389	43	15	145(37.3%)		3	41(10.5%)		12	104(26.7%)	
Immunology	CNT01959PSA3001	54160366	Arthritis, Psoriatic		46	20			5			15		
Immunology	CNT01959PSA3002	54160366	Arthritis, Psoriatic		89	30			8			22		
Immunology	CNT01959PSO3003	54160366	Psoriasis		87	38			10			28		
Immunology	CNT01959PSO3009	54160366	Psoriasis		124	42			15			27		
Neurosciences	54135419TRD3008	54135419	Treatment Resistant Depression	474	122	34	160(33.8%)		15	51(10.8%)		19	109(23.0%)	
Neurosciences	ESKETINTRD3001	54135419	Treatment Resistant Depression		45	21			7			14		
Neurosciences	ESKETINTRD3002	54135419	Treatment Resistant Depression		25	8			1			7		
Neurosciences	ESKETINTRD3003	54135419	Treatment Resistant Depression		80	27			9			18		
Neurosciences	ESKETINTRD3004	54135419	Treatment Resistant Depression		82	28			8			20		
Neurosciences	ESKETINTRD3005	54135419	Treatment Resistant Depression		17	4			2			2		
Neurosciences	R092670PSY3015	16977831	Psychosis		103	38			9			29		
Oncology	54179060CLL3011	54179060	Chronic Lymphocytic Leukemia	469	34	13	182(38.8%)		4	59(12.6%)		9	123(26.2%)	
Oncology	54767414AMY3001	54767414	Amyloidosis		49	22			5			17		
Oncology	54767414MMY3009	54767414	Cancer, Multiple Myeloma		18	10			4			6		
Oncology	54767414MMY3011	54767414	Cancer, Multiple Myeloma		15	7			4			3		
Oncology	56021927PCR3002	56021927	Metastatic Prostate Cancer		160	62			23			39		
Oncology	56021927PCR3003	56021927	Metastatic Prostate Cancer		173	58			17			41		
Oncology (2)	64091742PCR2001	64091742	Metastatic Prostate Cancer		20	10			2			8		
TOTAL					1418	528	37.2%		161	11.4%		367	25.9%	

Variables

Data – Full dataset (1418 sites)

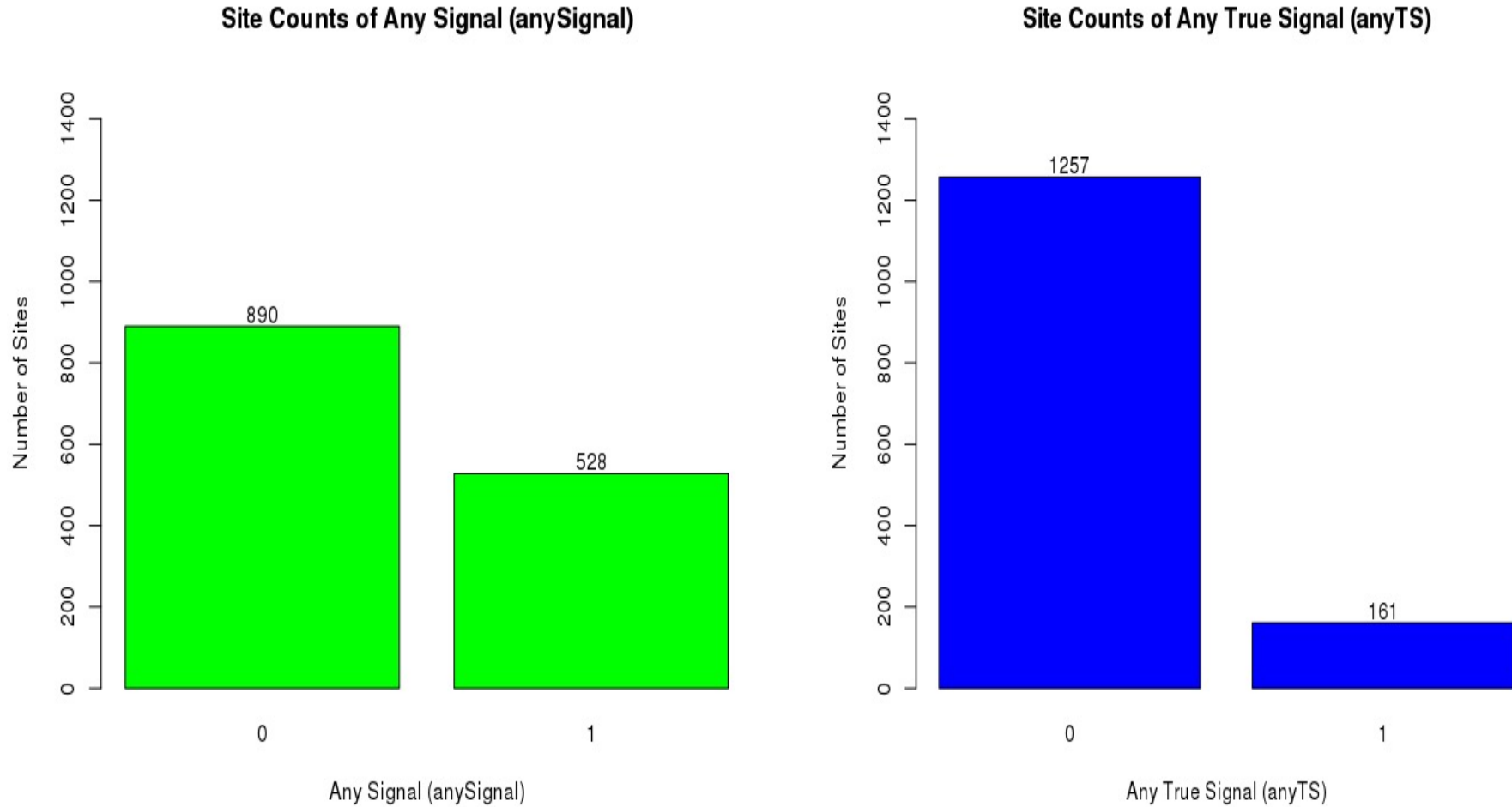
Response Variable

- Any True Signal

Predictor Variables (15)

- Number of subjects at that site for that study
- Compound
- Therapeutic Area
- Indication
- Number of patient weeks at that site for that study
- Country subregion
- EDC Domain
- Lab Domain
- Questionnaire Domain
- Intercept Statistic Test
- Slope Statistic Test
- Between Subject Statistic Test
- Within Subject Statistic Test
- Between Cluster Statistic Test
- Within Cluster Statistic Test

Signal Metrics – Any Signal & Any True Signal



Note: $367(528-161)$ sites have at least one signal and not a single True Signal(TS).

Summary of Preparation for Training Set

Ran 10+ different combinations of Predictor variables and 2 different Response variables (anySignal & anyTS) through GLM(exploring) & GBM(predicting) models

Final settings for model(GBM)

- Response = anyTS
- Predictors = default + 9 variables (EDC, LAB, QUEST, INT, SLOPE, B_SUB, W_SUB, B_CLUS & W_CLUS)

Split data into 70/30

- 70% (993 sites) to train model
- 30% (425 sites) to test model

Rebalanced the response due to class imbalance
(SMOTE R function)

- Unbalanced: anyTS=0 (880 sites) and anyTS=1 (113 sites)
- New: anyTS=0 (745 sites) and anyTS=1 (791 sites)

Ran GBM with Rebalanced Response anyTS for 10 folds and 10 repeats (100 models)

- AUC – Median (Range) – 99.3% (99.1% - 99.4%)
*Disclaimer: unrealistically high due to training dataset.
- Box plots (AUC) on later slides.
- Ordered variable importance on later slides.

Selected 1 out of the 100 models that performed good for the further predictions (e.g., test set)

Results - Test Set (425 sites)

- Model selected with prediction cut-off at 0.5 [range: 0-1]

- Confusion matrix

	FALSE			TRUE			Total
FALSE	344	91.3%	True Negative	33	8.8%	False Positive	377
TRUE	18	37.5%	False Negative	30	62.5%	True Positive	48
Total	362			63			425

- Accuracy → 88.0%
 - $(\text{true negative} + \text{true positive}) / \text{Total} = (344 + 30) / 425$
- AUC (Area Under the Curve) → 89.4%
 - Plot of true positive rate vs false positive rate
- Sensitivity → 62.5%
 - When we have a True Signal, are we predicting we have a True Signal?
 - $\text{true positive} / (\text{true positive} + \text{false negative}) = 30 / (30 + 18)$
- Specificity → 91.2%
 - When we have a False Signal, are we predicting we have a False Signal?
 - $\text{true negative} / (\text{true negative} + \text{false positive}) = 344 / (344 + 33) = 91.2\%$

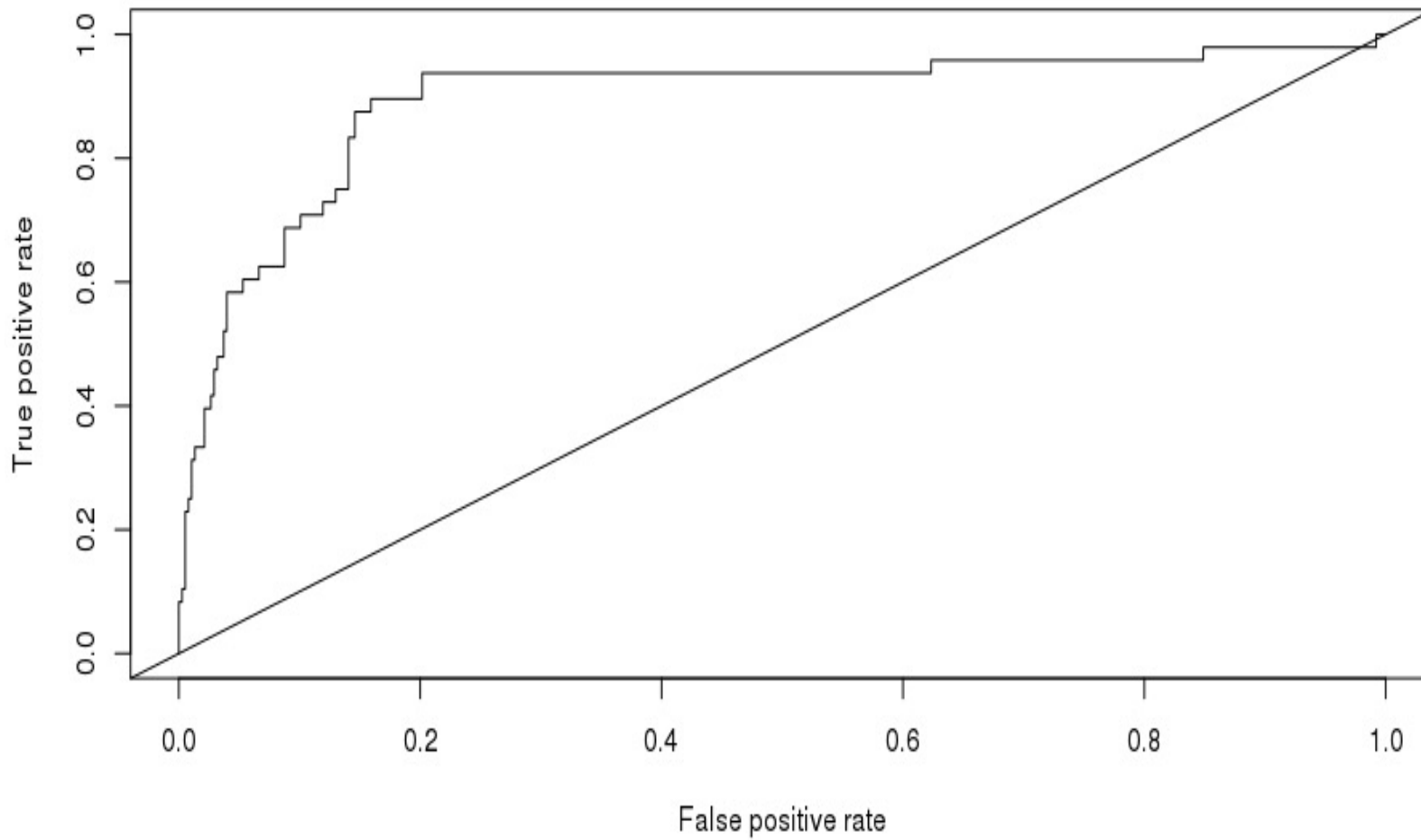
GBM Variable Importance

Variable Name	Relative Influence*
Questionnaire Domain	36.04
Within Subject Statistic Test	20.42
Within Cluster Statistic Test	7.60
EDC Domain	6.97
LAB Domain	5.44
Number of Patient Weeks at the Site	4.59
Intercept	4.57
Number of Subjects at the Site	3.75
Slope	2.73
Between Cluster Statistic Test	1.42
Between Subject Statistic Test	1.38
Subregion=Southern Europe	0.89

* Sum of all Relative Influence values will equal 100 (Normalized to 100).

AUC Plot - Test Set (425 sites)

AUC = 0.8943



Dashboard: TruePositive Predictions Tab

TruePositive Predictions

siteid	Totalreviews	# of Subjects	Therapeutic ...	Phase	Indication	# of Patient ...	Subregion	Probability	Priority for R...
UA10018	5	22	Immunology	2	CROHN'S DIS...	1385	Eastern Europe	1.00	True
PL10013	4	18	Immunology	2	CROHN'S DIS...	1234	Eastern Europe	0.97	True
GE10004	2	4	Immunology	2	CROHN'S DIS...	141	Western Asia	0.94	True
US10027	3	8	Immunology	2	CROHN'S DIS...	472	Northern Amer...	0.87	True
LV10002	2	6	Immunology	2	CROHN'S DIS...	502	Northern Europe	0.86	True
RS10004	2	6	Immunology	2	CROHN'S DIS...	280	Southern Europe	0.86	True
BA10005	1	3	Immunology	2	CROHN'S DIS...	172	Southern Europe	0.77	True
RS10005	1	9	Immunology	2	CROHN'S DIS...	600	Southern Europe	0.64	True
UA10005	1	4	Immunology	2	CROHN'S DIS...	308	Eastern Europe	0.46	False
SK10001	1	5	Immunology	2	CROHN'S DIS...	199	Eastern Europe	0.46	False
TV10001	1	5	Immunology	2	CROHN'S DIS...	312	Eastern Asia	0.38	False
US10080	1	3	Immunology	2	CROHN'S DIS...	271	Northern Amer...	0.33	False
US10023	2	8	Immunology	2	CROHN'S DIS...	686	Northern Amer...	0.28	False
UA10009	2	37	Immunology	2	CROHN'S DIS...	1754	Eastern Europe	0.26	False
PL10017	1	9	Immunology	2	CROHN'S DIS...	847	Eastern Europe	0.26	False
US10064	4	4	Immunology	2	CROHN'S DIS...	567	Northern Amer...	0.25	False
DE10016	1	7	Immunology	2	CROHN'S DIS...	401	Western Europe	0.21	False
IT10013	1	4	Immunology	2	CROHN'S DIS...	355	Southern Europe	0.18	False
US10012	1	10	Immunology	2	CROHN'S DIS...	573	Northern Amer...	0.16	False
IL10009	1	8	Immunology	2	CROHN'S DIS...	336	Western Asia	0.12	False
PL10026	1	10	Immunology	2	CROHN'S DIS...	711	Eastern Europe	0.02	False
PL10002	1	21	Immunology	2	CROHN'S DIS...	1337	Eastern Europe	0.02	False
PL10025	0	16	Immunology	2	CROHN'S DIS...	1023	Eastern Europe	0.01	False
PL10010	1	11	Immunology	2	CROHN'S DIS...	1052	Eastern Europe	0.01	False
MY10002	0	6	Immunology	2	CROHN'S DIS...	175	South-Eastern ...	0.01	False
DE10007	0	3	Immunology	2	CROHN'S DIS...	148	Western Europe	0.01	False
PL10005	0	3	Immunology	2	CROHN'S DIS...	160	Eastern Europe	0.01	False
PL10021	0	4	Immunology	2	CROHN'S DIS...	256	Eastern Europe	0.01	False
CA10003	0	3	Immunology	2	CROHN'S DIS...	55	Northern Amer...	0.01	False
US10011	0	3	Immunology	2	CROHN'S DIS...	276	Northern Amer...	0.01	False
JO10002	0	7	Immunology	2	CROHN'S DIS...	298	Western Asia	0.01	False
CZ10004	0	3	Immunology	2	CROHN'S DIS...	133	Eastern Europe	0.01	False
SK10002	0	5	Immunology	2	CROHN'S DIS...	310	Eastern Europe	0.01	False
FR10010	0	3	Immunology	2	CROHN'S DIS...	213	Western Europe	0.01	False
KR10011	0	4	Immunology	2	CROHN'S DIS...	126	Eastern Asia	0.01	False
NL10004	0	6	Immunology	2	CROHN'S DIS...	126	Western Europe	0.01	False
RU10018	0	4	Immunology	2	CROHN'S DIS...	117	Eastern Europe	0.01	False
PL10001	0	11	Immunology	2	CROHN'S DIS...	1061	Eastern Europe	0.01	False
UA10021	0	4	Immunology	2	CROHN'S DIS...	64	Eastern Europe	0.01	False
RS10002	0	8	Immunology	2	CROHN'S DIS...	332	Southern Europe	0.01	False
IL10001	0	3	Immunology	2	CROHN'S DIS...	63	Western Asia	0.01	False
HU10005	0	4	Immunology	2	CROHN'S DIS...	70	Eastern Europe	0.01	False
JO10005	0	5	Immunology	2	CROHN'S DIS...	84	Western Asia	0.00	False
RU10014	0	8	Immunology	2	CROHN'S DIS...	294	Eastern Europe	0.00	False
AU10008	0	3	Immunology	2	CROHN'S DIS...	85	Australia and ...	0.00	False
RU10020	0	9	Immunology	2	CROHN'S DIS...	425	Eastern Europe	0.00	False
PL10018	0	11	Immunology	2	CROHN'S DIS...	724	Eastern Europe	0.00	False
US10057	0	3	Immunology	2	CROHN'S DIS...	238	Northern Amer...	0.00	False
RU10029	0	4	Immunology	2	CROHN'S DIS...	38	Eastern Europe	0.00	False
BY10001	0	5	Immunology	2	CROHN'S DIS...	231	Eastern Europe	0.00	False
PL10011	0	4	Immunology	2	CROHN'S DIS...	230	Eastern Europe	0.00	False
AT10003	0	3	Immunology	2	CROHN'S DIS...	231	Western Europe	0.00	False
RU10008	0	13	Immunology	2	CROHN'S DIS...	569	Eastern Europe	0.00	False

Filters

Type to search filters

singleStudyTPPrediction

Priority for Review

☒ False

☒ True

totalreviews

0 5

Probability

0.00 1.00

of Subjects

3 37

Indication

☒ CROHN'S DISEASE

of Patient Weeks

38 1754

Subregion

☒ Australia and New Zealand

☒ Eastern Asia

☒ Eastern Europe

☒ Northern America

☒ Northern Europe

☒ South-Eastern Asia

☒ Southern Europe

☒ Western Asia

☒ Western Europe

siteid

Type to search in list

(All) 53 values

AT10003

AU10008

BA10005

BY10001

CA10003

CZ10004

Therapeutic Area

☒ Immunology

Phase

☒ 2

53 of 53 rows 0 marked 10 columns singleStudyTPPrediction

Conclusions

- Decreased false positives by $\sim 65\%$.
- Reduced amount of time to analysis a study significantly.
- Or implement focused approach:
 - First analysts focus on the sites with a positive signal prediction.
 - Then, analysts can proceed to investigate all remaining signals or pick & choose which other signals they feel warrant investigating.
- By continuing to add more studies and sites to the model, predication probabilities will become more accurate.



PHARMACEUTICAL COMPANIES OF
Johnson & Johnson