



Implementing Advanced Analytics and Machine Learning in Risk Based Strategies

Dr Rajneesh Patil BDS., MBA., FPH
Head: Process, Technology, Analytics
Clinical Operations - IQVIA

Talking Points

- + A view of the industry trend in healthcare analytics
- + Risk Based Monitoring (RBM) - Model
- + Analytics capabilities in executing risk based strategies
- + Case studies in implementation of machine learning models
- + Highlights and conclusory remarks

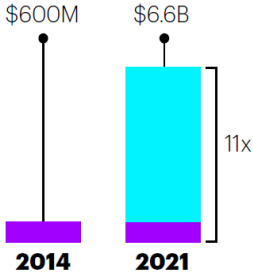
Analytics in healthcare - Survey

FIGURE 1.

The AI health market is seeing explosive growth

HEALTH AI MARKET SIZE 2014-2021

Acquisitions of AI startups are rapidly increasing while the health AI market is set to register an explosive CAGR of 40% through 2021



Source: Accenture analysis

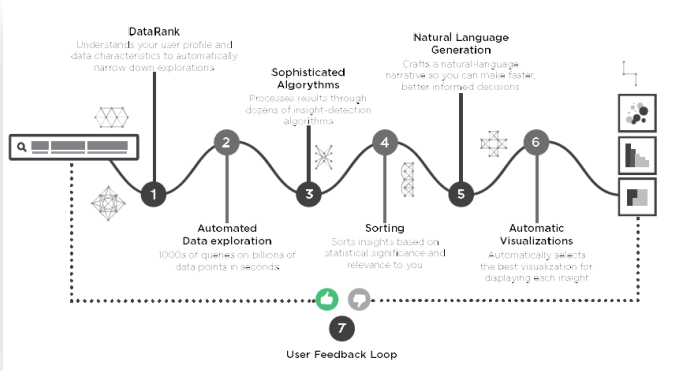
Top 10 AI Applications

APPLICATION	VALUE*	APPLICATION	VALUE*
Robot-Assisted Surgery**	\$40B	Connected Machines	\$14B
Virtual Nursing Assistants	\$20B	Clinical Trial Participant Identifier	\$13B
Administrative Workflow Assistance	\$18B	Preliminary Diagnosis	\$5B
Fraud Detection	\$17B	Automated Image Diagnosis	\$3B
Dosage Error Reduction	\$16B	Cybersecurity	\$2B
		TOTAL = ~\$150B	

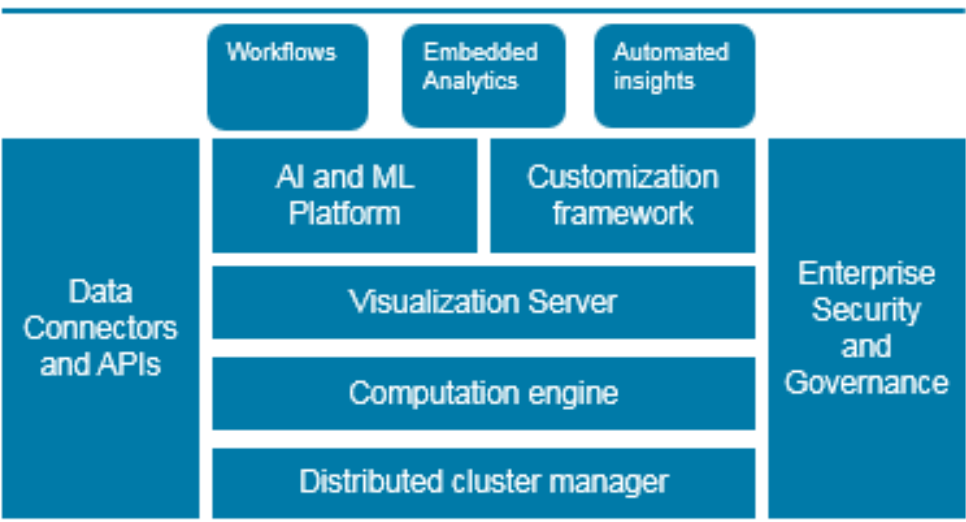
● Areas we already touch through our initiatives

ML

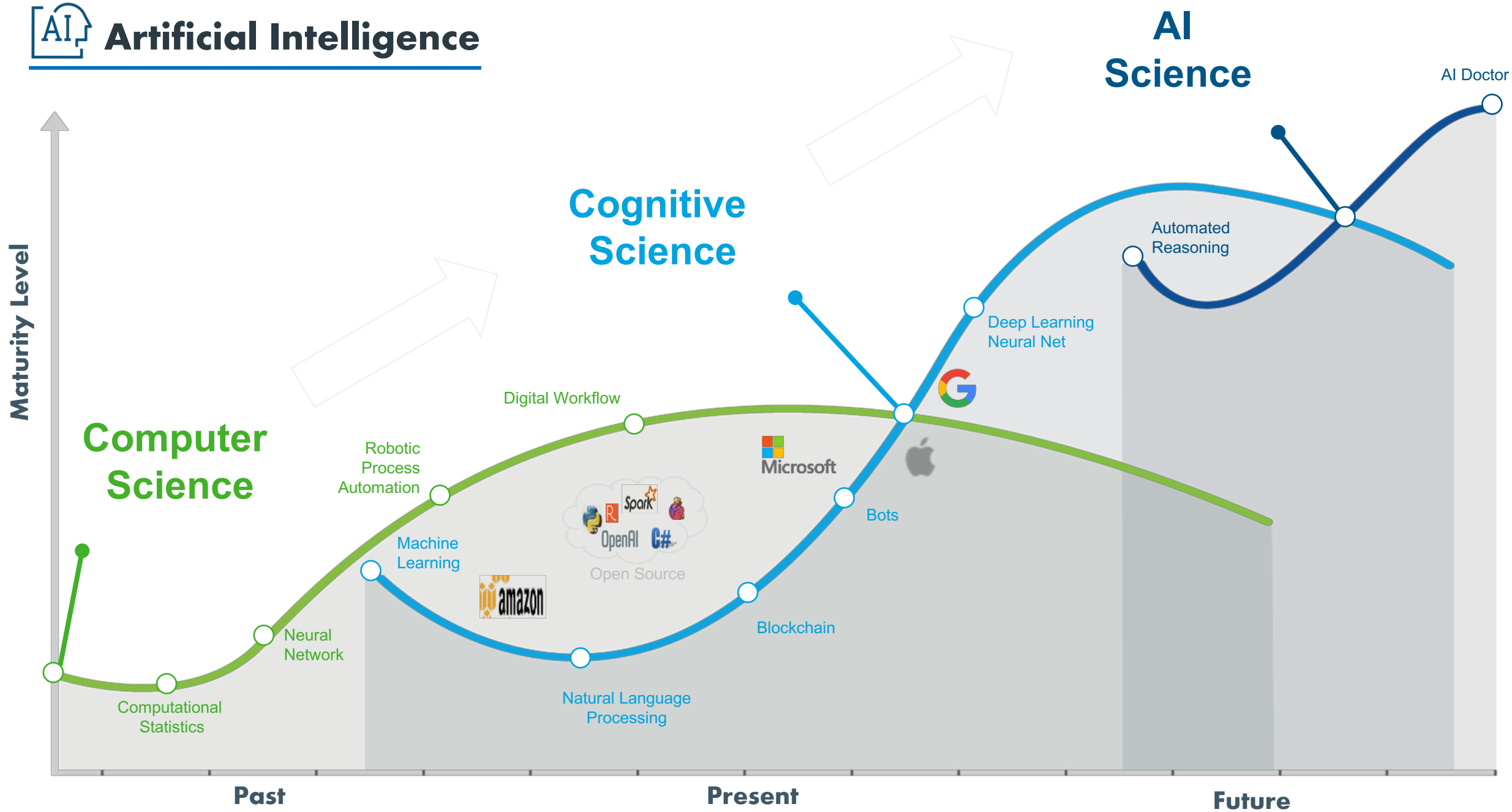
AI



DATA APPS

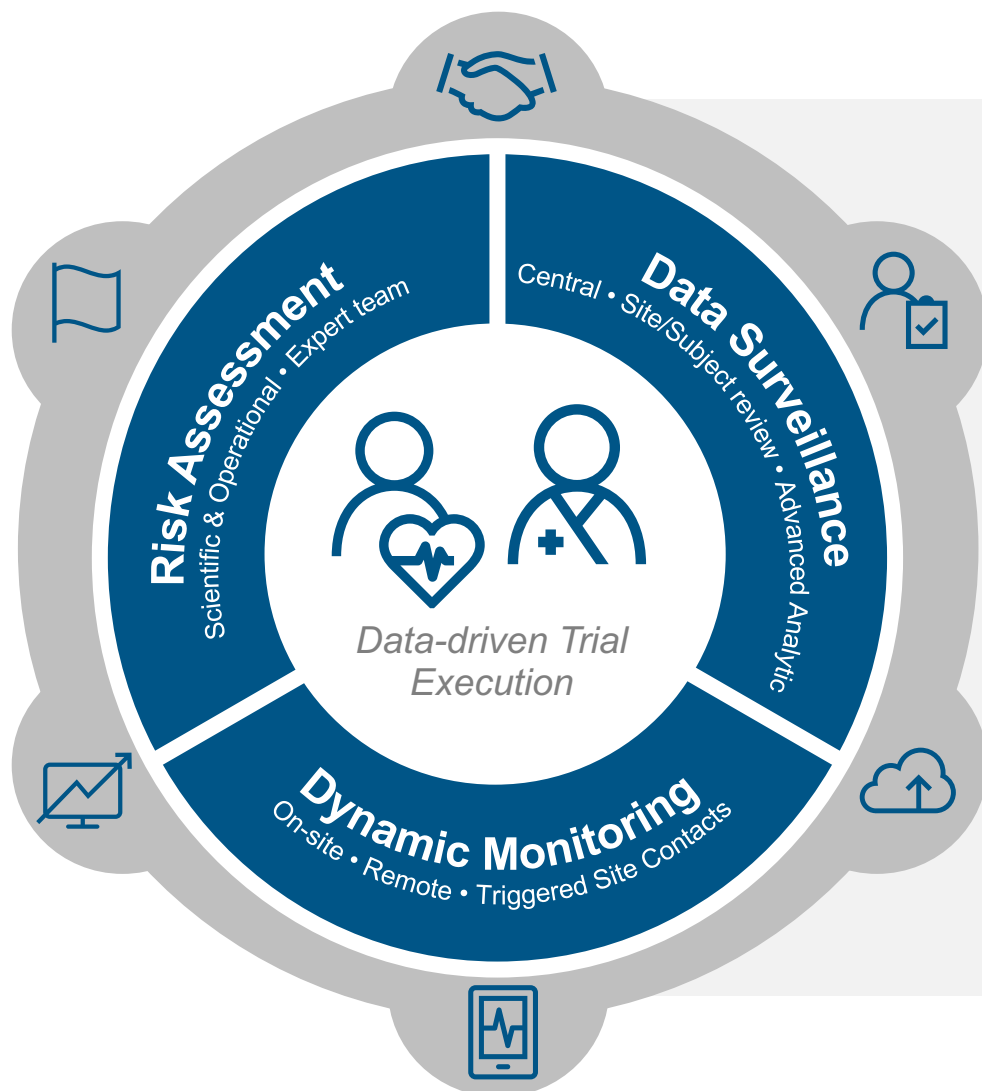


* Source: MIT, Accenture : Trends in Healthcare technology investment



Data-driven Trial Execution

Identify & mitigate risk, use the right monitoring approach, identify potential issues faster



- **Risk Assessment Mitigation Plan (RAMP):**
 - IQVIA and sponsor jointly perform the study risk assessment
 - Identification of critical variables and processes and optimal monitoring strategy
- **Dynamic Monitoring:**
 - *On-Site Monitoring:* Source Data Review, Source Data Verification, ICF review, IP Accountability, File Review
 - *Remote Monitoring:* Data entry, Data cleaning, Enrollment discussion, Event capture and documentation
 - *Centralized Monitoring:* Holistic 'real time' patient CRF and critical variables
- **Data surveillance:**
 - Site & patient level review and advanced analytics to identify trends and outliers

Implementing Centralized Monitoring Analytics

Implementing analytics models across advanced, predictive and machine learning



**Advanced
Analytics**

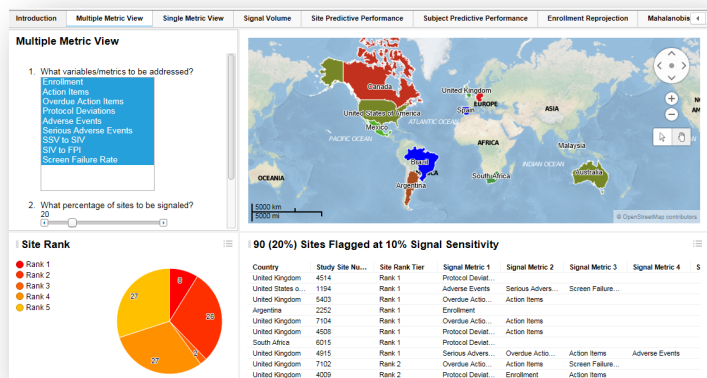


**Predictive
Analytics**

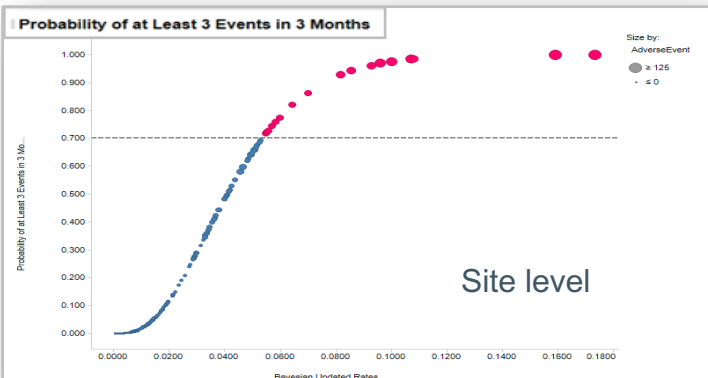


**Machine
Learning**

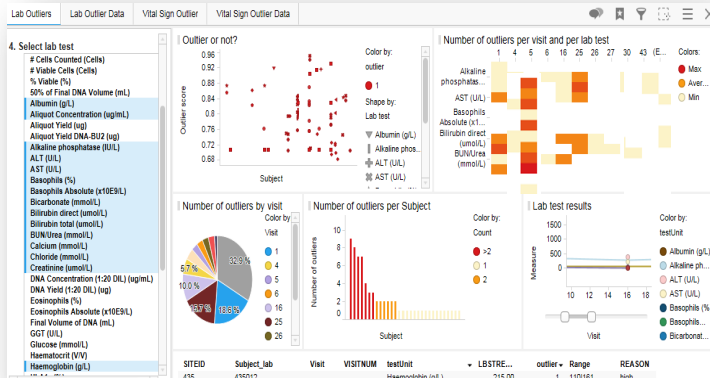
Composite site ranking for holistic risk assessment at sites



Which sites are at high risk of protocol non-compliance?



Which subjects are at potential safety risk with high outliers in lab analytes



Case study 1

Predictive modelling of AE reporting risks through machine learning methodology

Feature selection

Primary response feature: “HR_Site” (High Risk Site, in which "Y" corresponds to any site with unreported AE's).

Secondary response feature: “Unreported_AE's_perSite” (Number of Unreported AEs per Site)

Effects of scaling features: Scaled features (by study) did not improve the overall model prediction accuracy noticeably, and they were not listed as the most important features, so only the original features were used in the further analyses

Handle multicollinearity: 1) *Pearson's* $r \geq 0.9$ as threshold to remove correlated features; 2) bagging and regularization algorithms in models

Selected predictor features:

- Avg_AEReportingLag_perSite (also Median, and SD)
- Avg_AEs_per_ReportingEvent_perSite (also Median, and SD)
- Avg_Difference_PlannedtoActual_VisitDate_perSite (also Median, and SD)
- Avg_Pt_Time_onStudy_perSite
- Avg_Query_ResponseTime_perSite (also Median, and SD)
- Avg_QueryIssue_toClose_perSite (also Median, and SD)
- Avg_VisitEntryLag_perSite (also Median, and SD)
- DataPages_perSite
- FollowUp_Visits_perSite
- Number_AEs_withTimetoCCOD_LessThan_SiteAvgLag_perSite
- Number_Visits_withTimetoCCOD_LessThan_SiteAvgLag_perSite
- Pts_withVisits_perSite
- QueryRate_perSite
- Reported_AEs_perSite
- SeriousAEs_perSite
- Unanswered_Queries_perSite
- Unreported_AEs_perSite
- Unscheduled_Visits_perSite
- Visits_perSite
- Country
- Indication
- Region
- Sourcing
- Therapeutic Area

Develop predictive algorithms

Types of models implemented and compared:

- Random Forest (RF)
- Gradient Boosting Trees (GBM)
- Deep Neural Network (DNN)
- Logistic Model with Regularization (GLM)

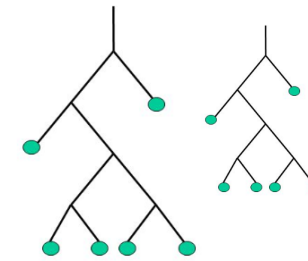
Tune important hyper parameters (e.g., GBM model):

- Number of trees
- Learning rate
- Learn rate annealing
- Max depth of tree
- Min improvement at each split

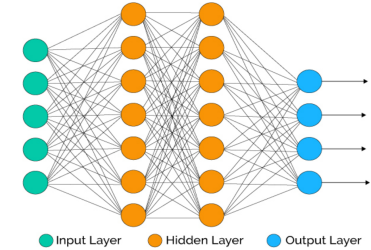
Develop a series of models with sampled parameters:

- Compare AUC (Area under the ROC curve) of models in the test data
- Store parameters of 'best' model with the highest AUC

Machine-learning algorithms

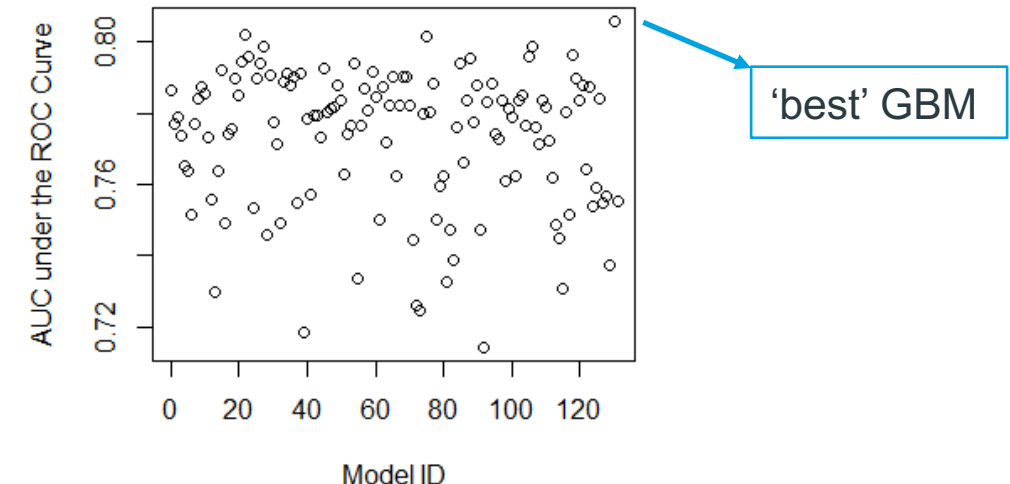


Tree-based model



Neural Network

Tuning parameters based on grid search



Model Evaluation and Validation

Metrics for model evaluation and validation:

AUC: Area under Receiver-Operating-Characteristic curve

Accuracy: proportion of correctly classified observations

Specificity: True negative/condition negative

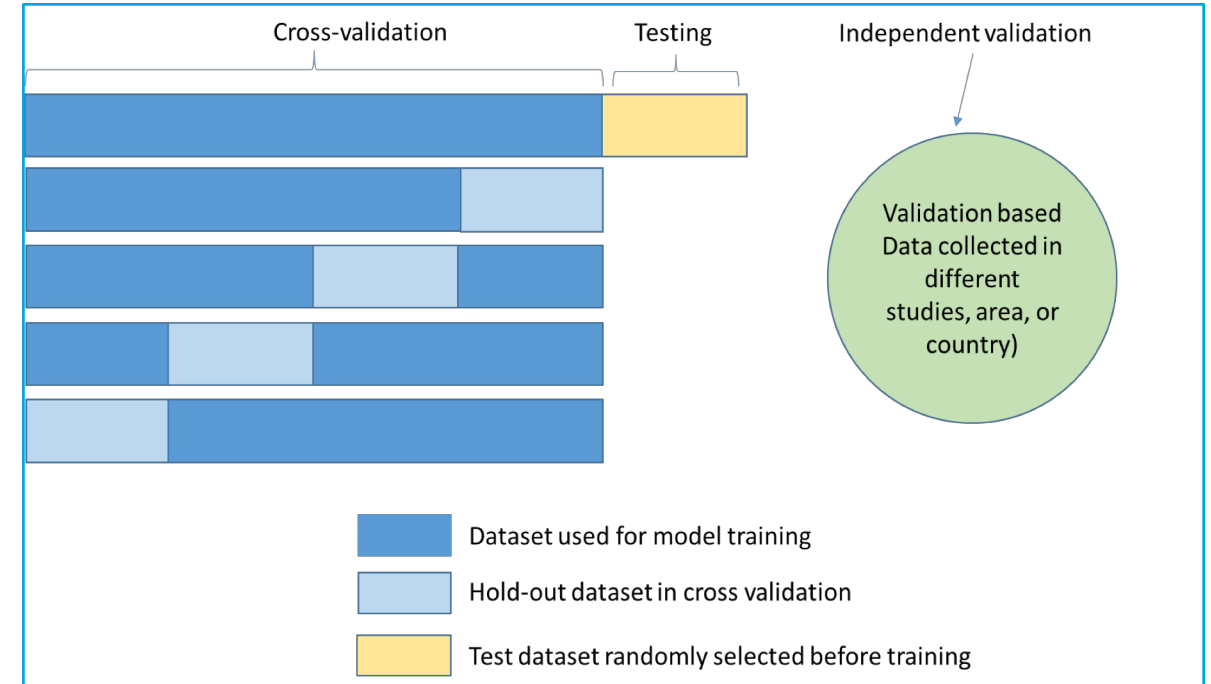
Recall: True positive/condition positive

Precision: True positive/predicted condition positive

F1: $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$

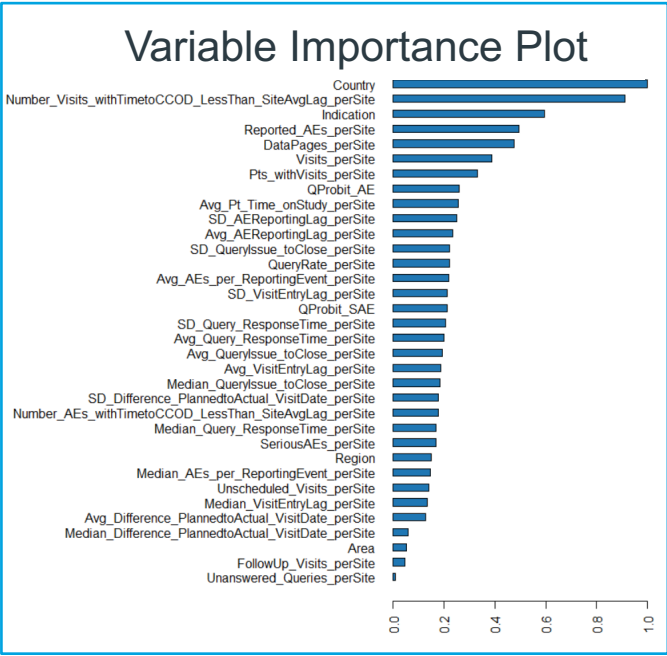
Evaluation and validation were performed in following processes:

- 1) Model training
- 2) X-fold Cross-Validation based on the training data
- 3) Model testing (step 1: Randomly sample 20% of data from the raw data as the test data before model training; and step 2: Evaluate model performance based on this test data)

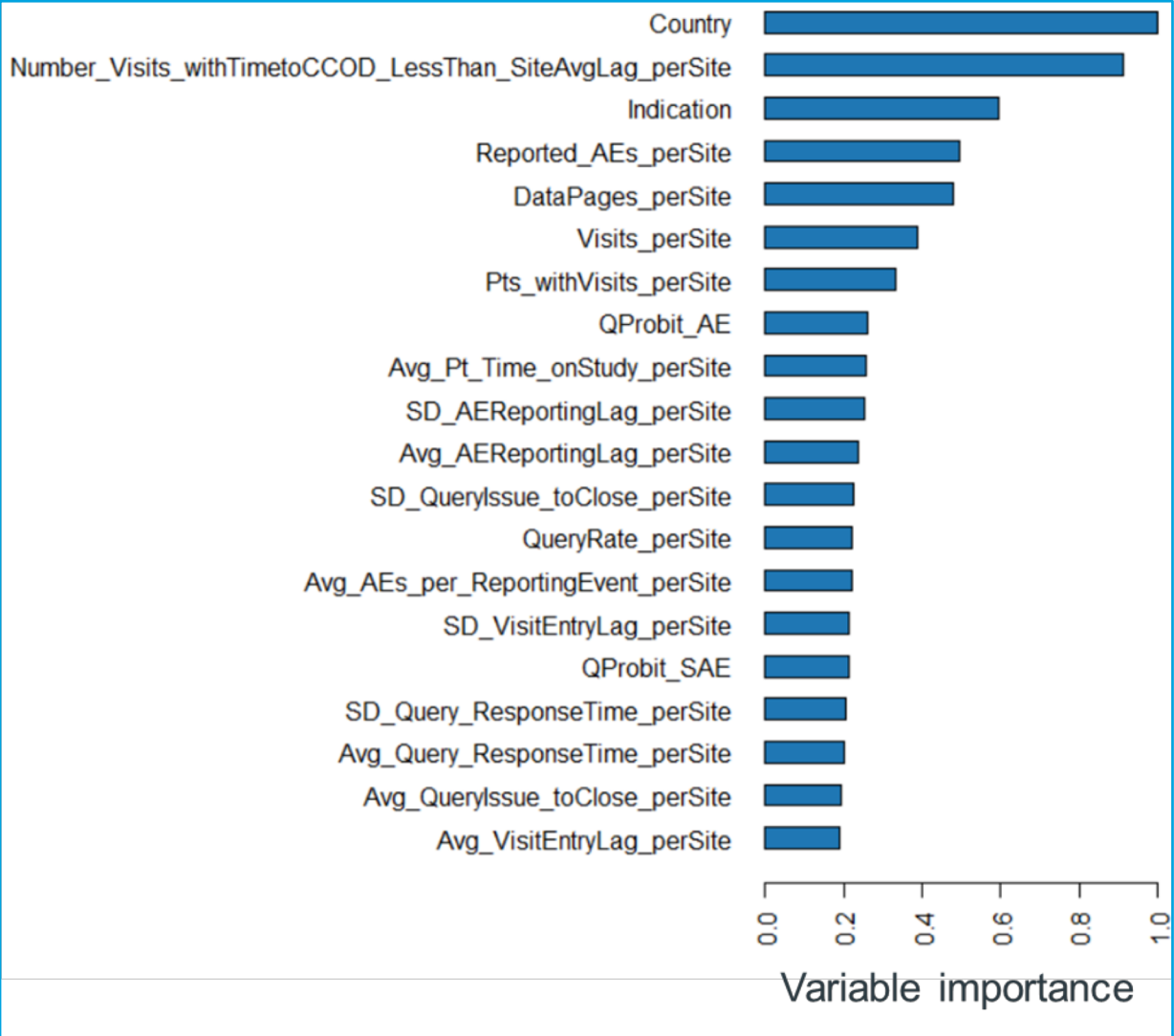


Identifying Key Risk Indicators

Ranking the importance (or contribution to behavior) of predictor features in the **RF** model for predicting “High-Risk Sites” with unreported AEs

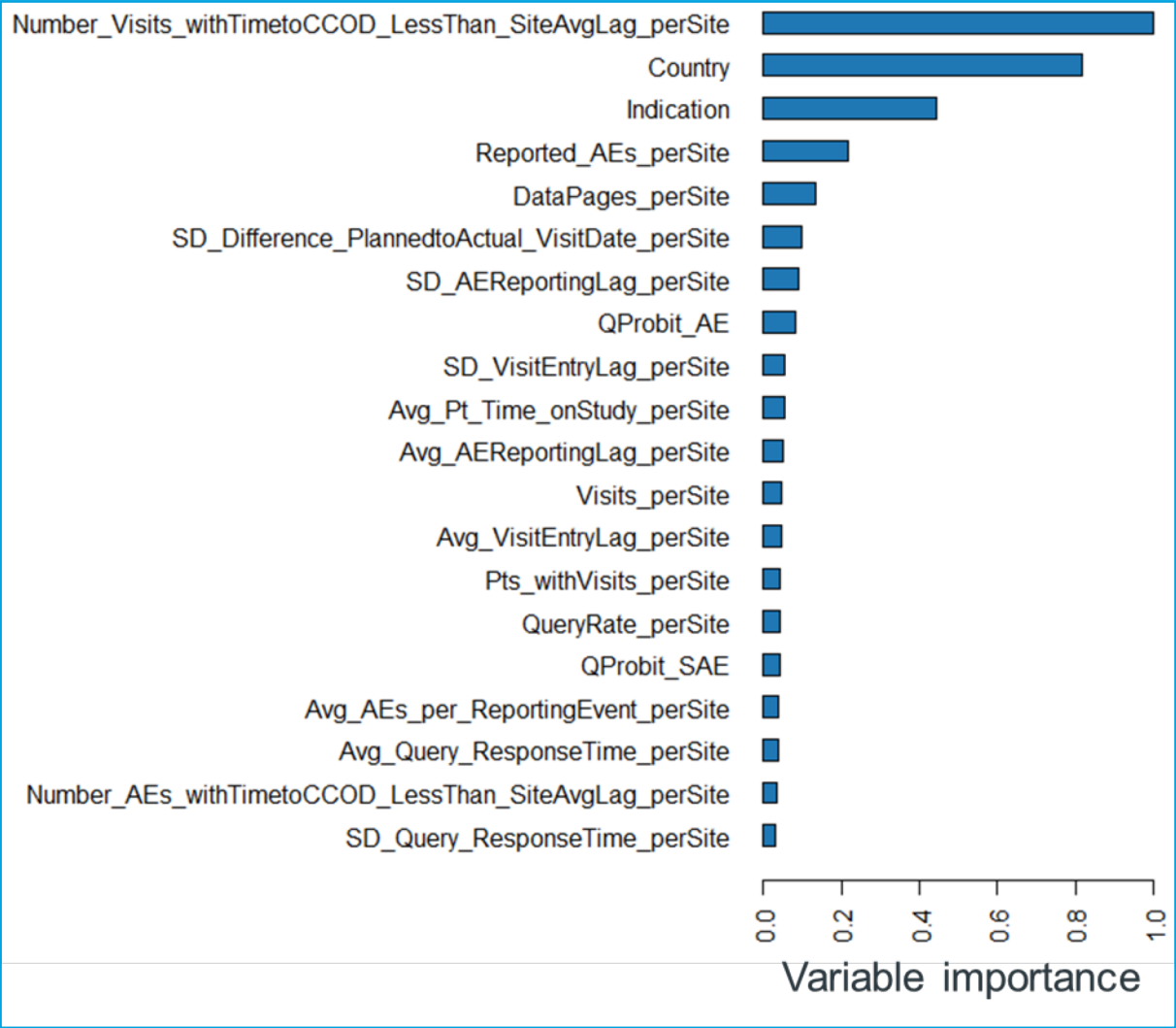
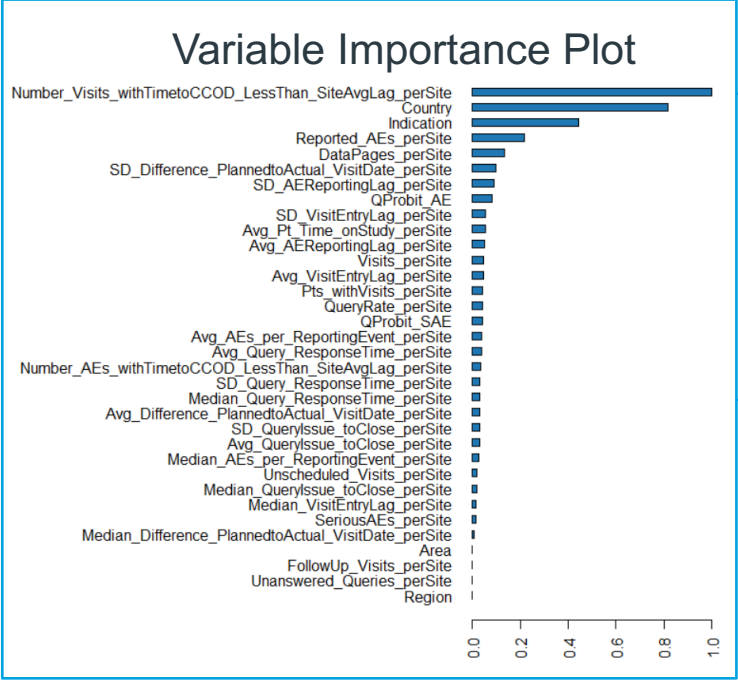


Top 20



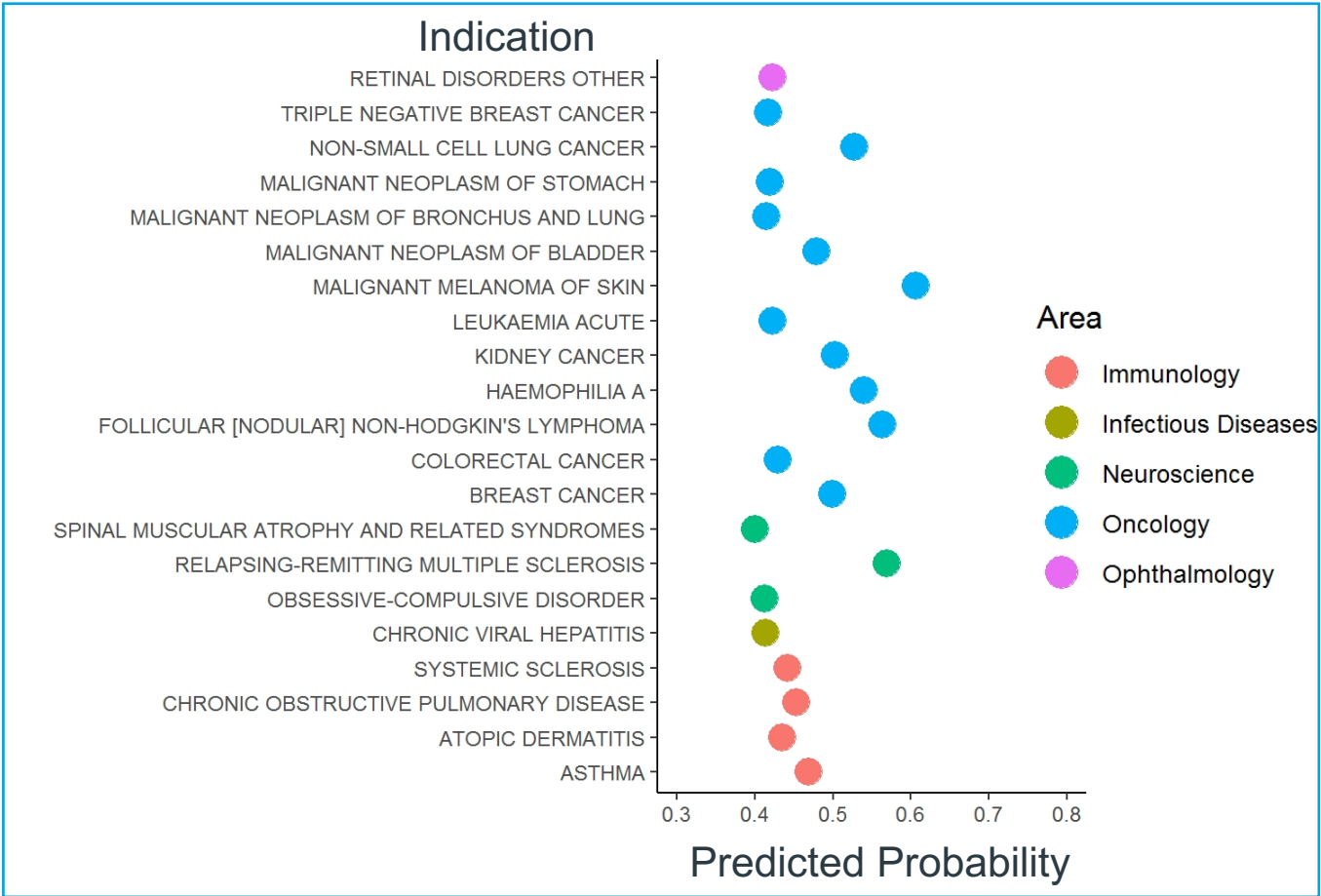
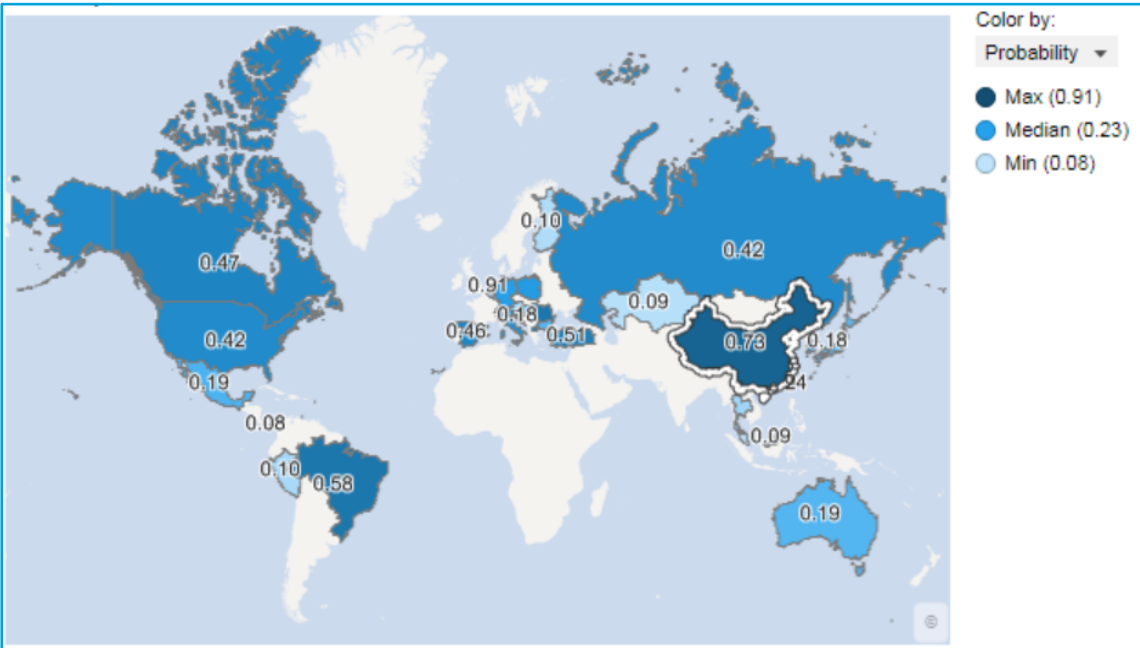
Identifying Key Risk Indicators

Ranking the importance (or contribution to behavior) of predictor features in the **GBM** model for predicting “High-Risk Sites” with unreported AEs



Effect of Key Risk Indicators

Geographical heterogeneity in predicted probability of risk (not/under/delayed reporting of AEs)

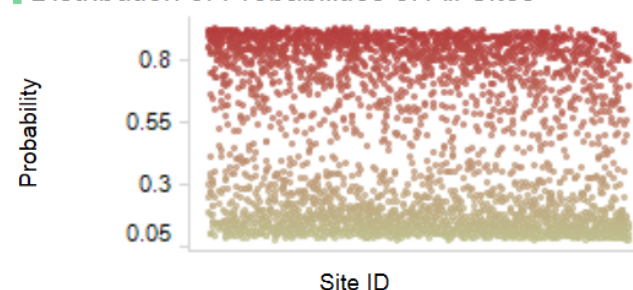


Partial dependence plot: Effect of indication on probability of risk marginalizing over all other features.

Visualization: Predictive Analytics

Predict probability of not/under/delayed reporting of AEs at each site

Distribution of Probabilities of All Sites



Validate model performance using misclassification table

Predicted Vs Observed High-Risk Sites
1-Yes; 0-No

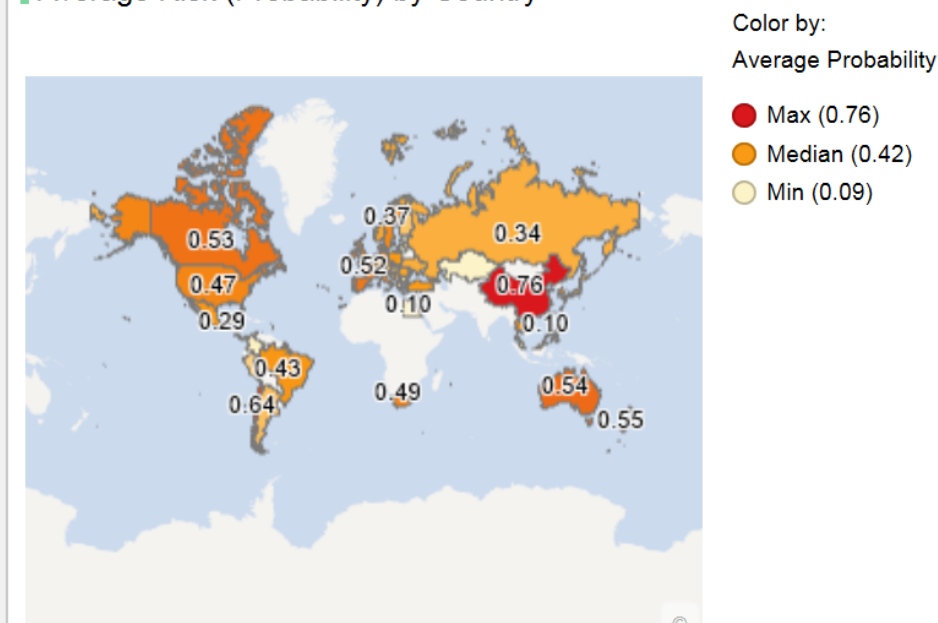
		Predicted	
		0	1
Observed	0	1821	92
	1	102	1481

Evaluate Key Risk Indicators for selected or signaled sites

Key Risk Indicators

Probability	Country	Indication	Serious AEs	Pts_with Visits	Data Pag	Number_Visits_withTimetoCCOD_LessThan_SiteAvgLag
0.63	UNITED S...	HAEMO...	0	3	59	7
0.37	UNITED K...	HAEMO...	1	2	26	6
0.81	UNITED S...	HAEMO...	1	3	65	13
0.79	UNITED S...	HAEMO...	0	2	26	2
0.84	UNITED S...	HAEMO...	2	7	104	10
0.85	POLAND	HAEMO...	4	6	81	23

Average Risk (Probability) by Country



Color by:
Average Probability

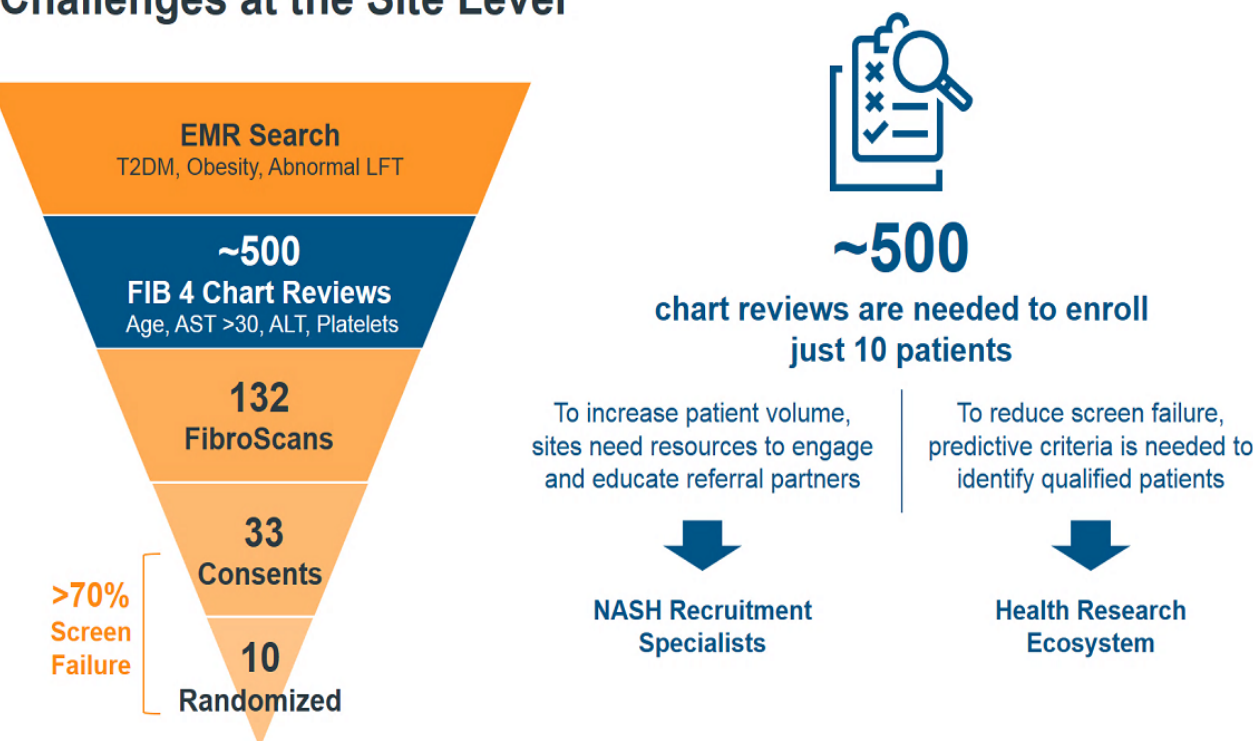
- Max (0.76)
- Median (0.42)
- Min (0.09)

Case Study 2

Early diagnosis of:
Nonalcoholic Steatohepatitis

Challenges and Project Objectives

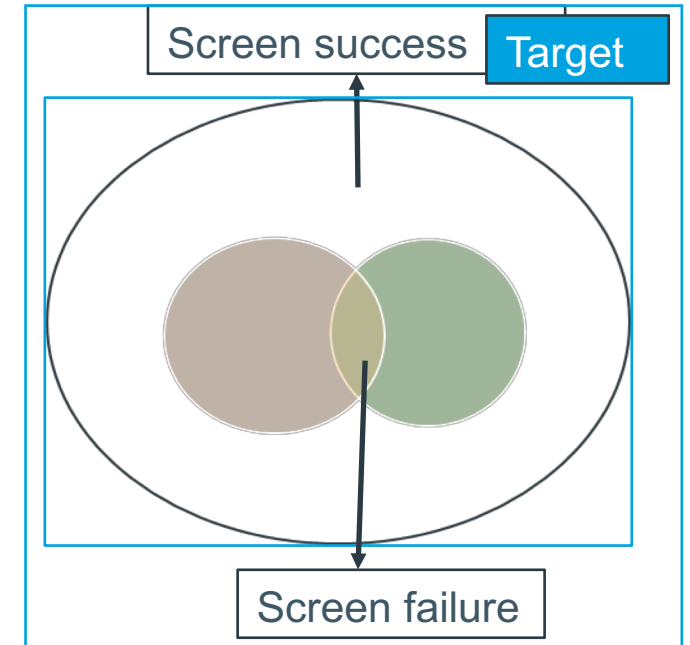
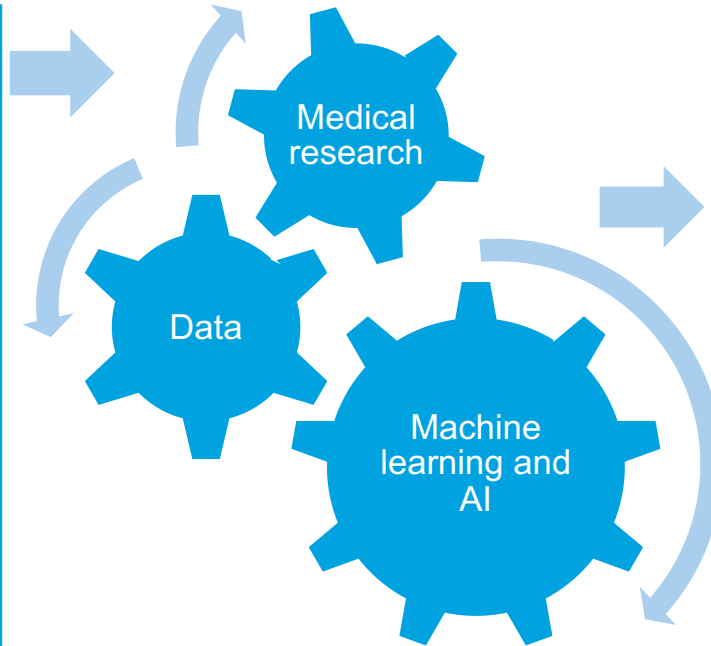
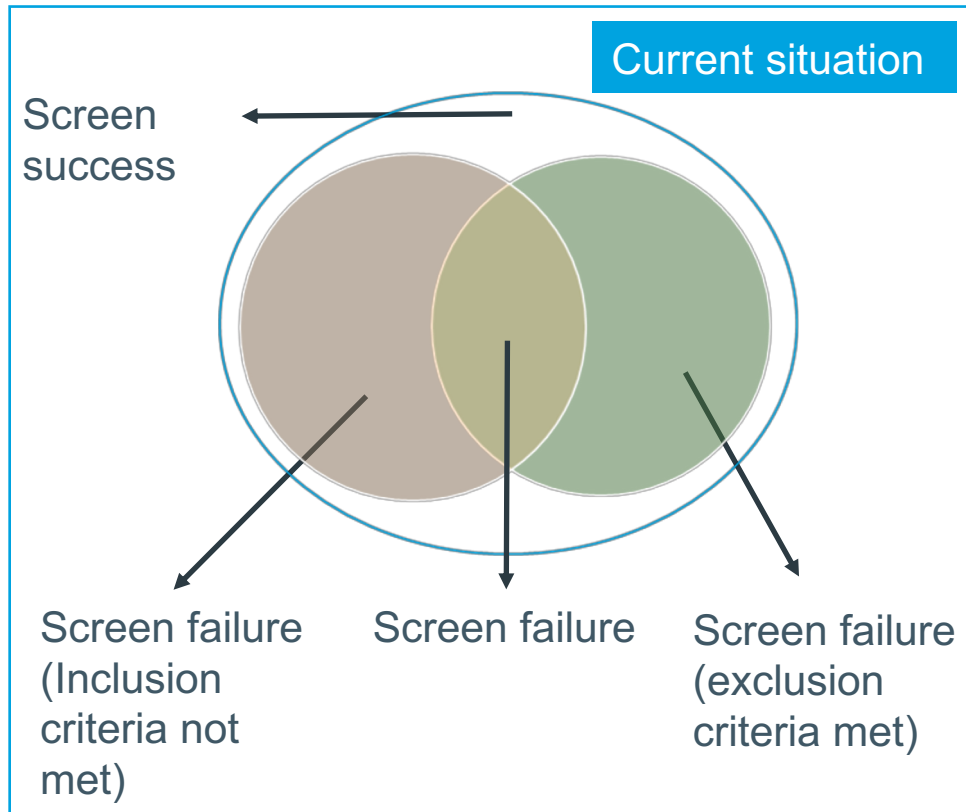
Challenges at the Site Level



Objectives:

- Implement statistical and machine learning algorithms to detect undiagnosed NASH patients
- Develop an artificial intelligence framework to automate early detection of NASH patients
- Test the feasibility of Ambulatory EMR database
- Understand overall geographic distribution of NASH patients

Proposed Framework



- Non-invasive diagnosis of NASH (liver biopsy is optional)
 - Scalable to other database with similar structure as EMR data
 - Automatic weights of risk factors by machine learning models
 - Versatile to account from both linear and non-linear interactions
 - Probability-based, allowing tradeoff between sensitivity and specificity, and between recall and precision
- Advantages**

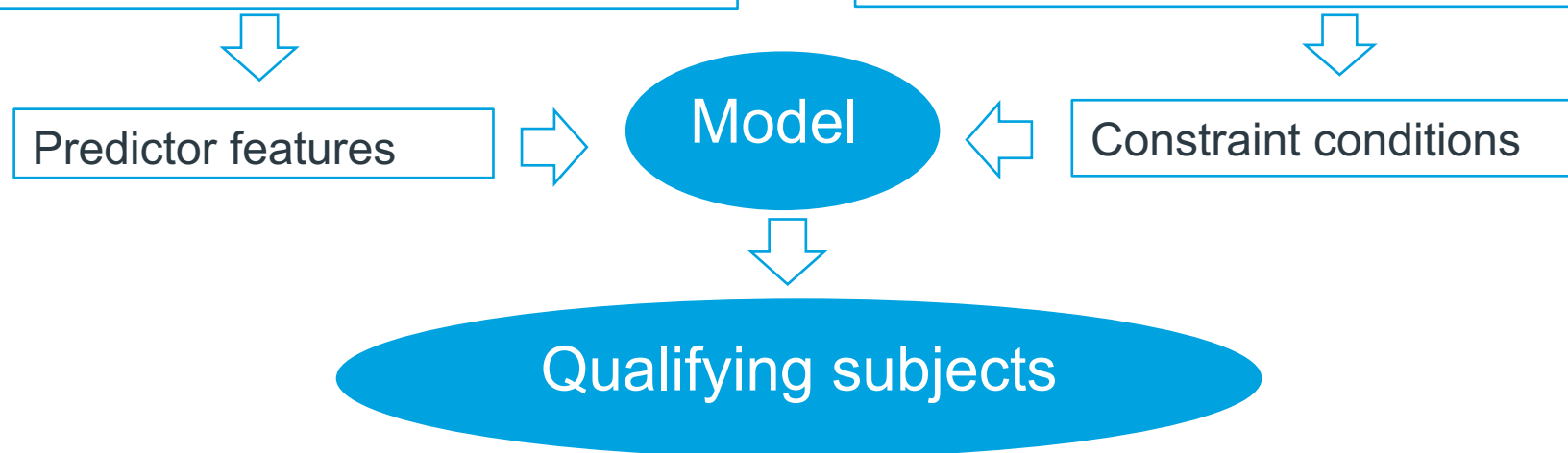
Inclusion & Exclusion Assessment

NASH Risk Factors:

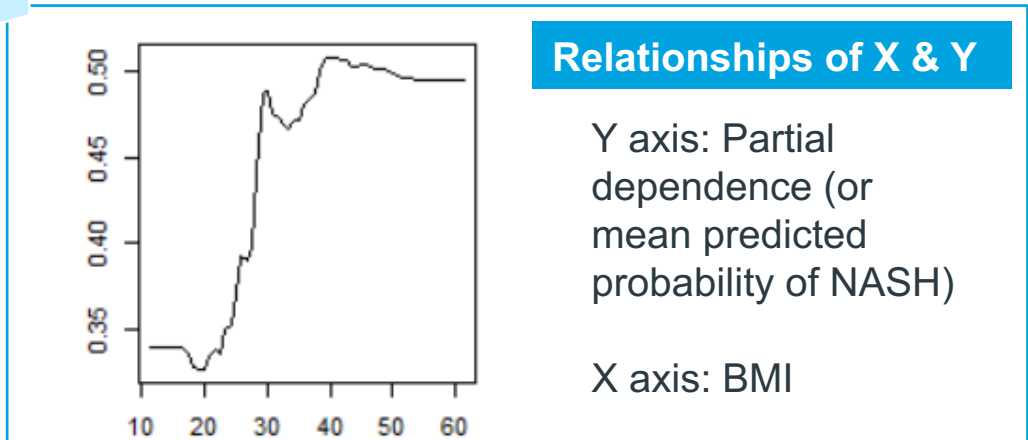
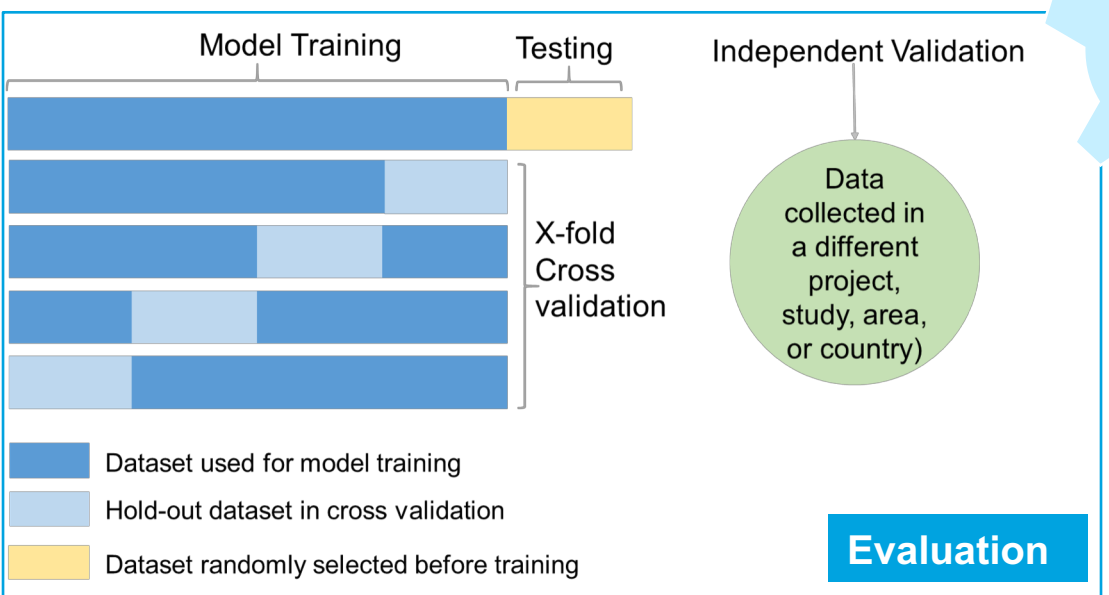
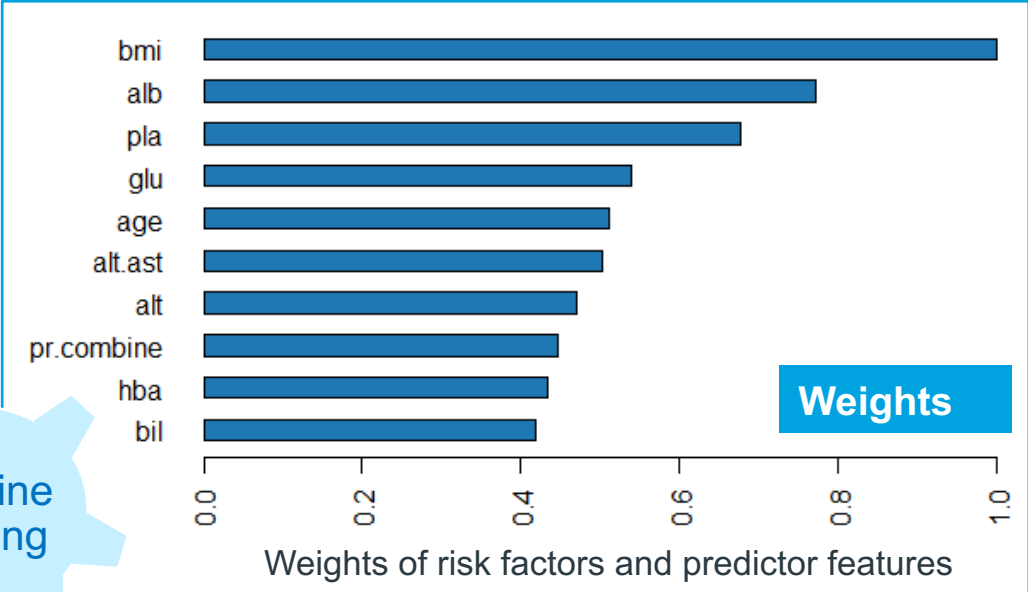
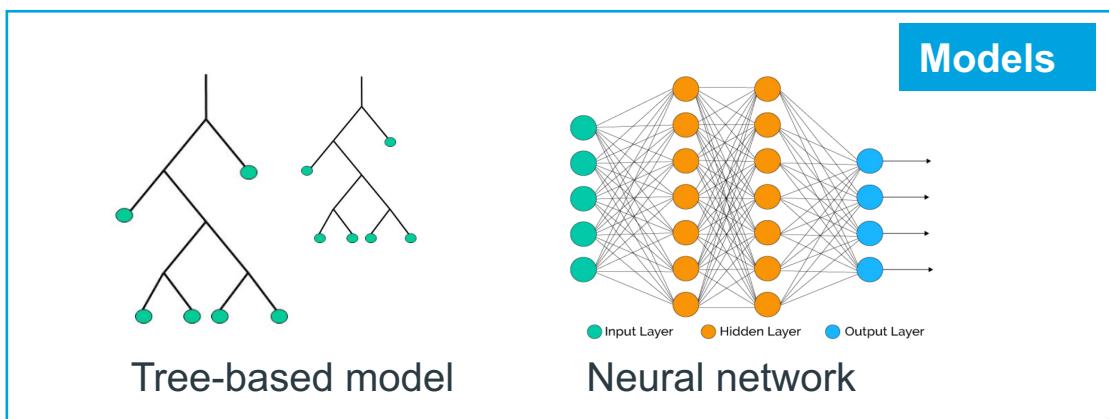
- Type 2 DM
- ...
- Any medication to treat type 2 DM
- Any medication prescribed for weight loss
- ...
- Polycystic ovary disease

Common Exclusion Criteria in NASH:

- Congestive heart failure
- ...
- Drug abuse
- Alcohol abuse
- ...
- Other liver diseases



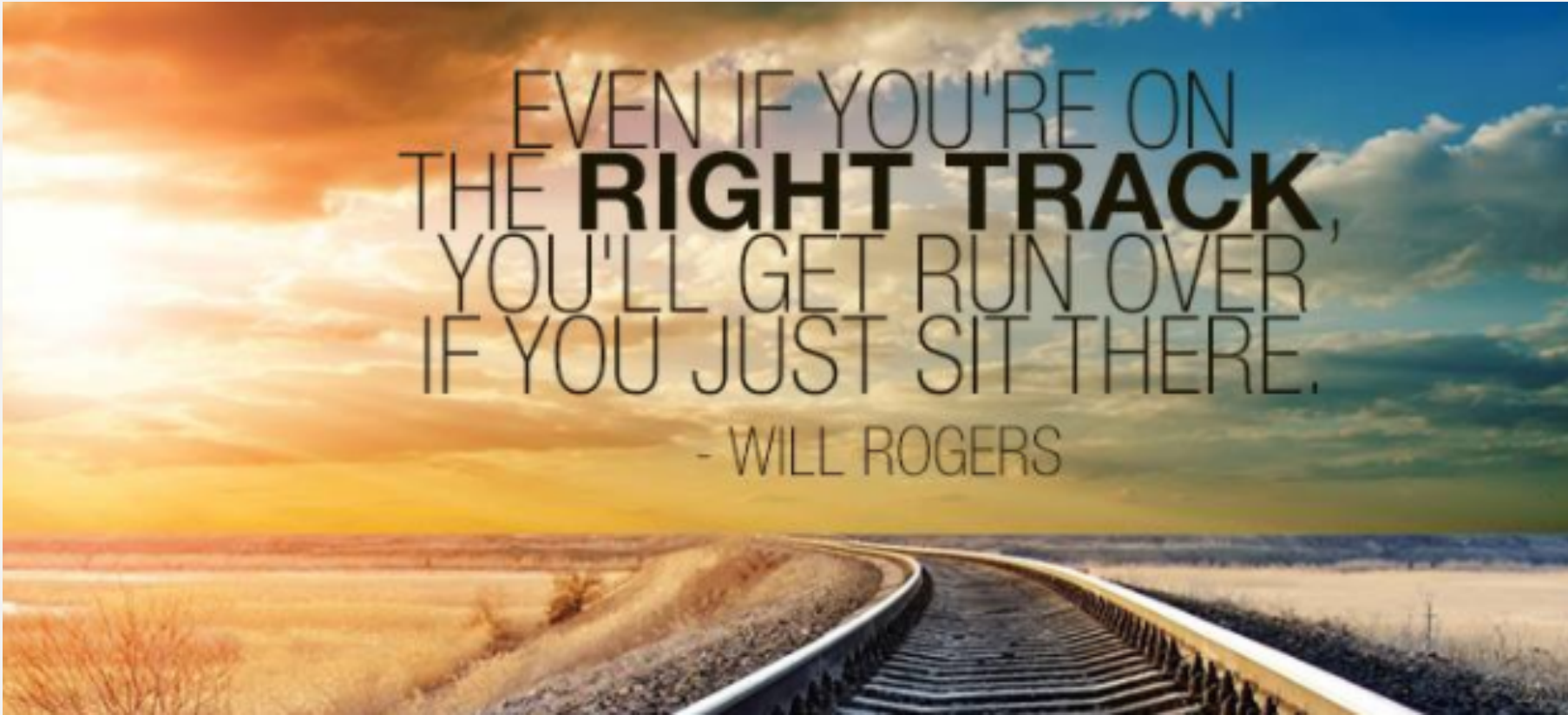
Develop Predictive Models



Closing Remarks

Highlights

- 
- Use of advanced and contemporary statistical models can provide composite assessment of site risk score by combining several key risk indicators into one comprehensive risk score.
 - Helps reduce the white noise/false positives in terms of site risk identification (compared to traditional statistical models)
 - Use of machine learning models in subject outlier detection, statistical monitoring of patient data and predictive modelling to identify sites at risk in factors like delayed or underreporting.
 - Other applications in early diagnosis of potential rare disease indications to facilitate patient identification and site selection strategies.



EVEN IF YOU'RE ON
THE **RIGHT TRACK**,
YOU'LL GET RUN OVER
IF YOU JUST SIT THERE.
- WILL ROGERS