Experiences from running internal prediction challenges within a pharmaceutical company

BBS Seminar : Predictive modelling, machine learning and causality

November 1st 2019

Chris Harbron
Expert Statistical Scientist, Roche Biostatistics MCO

Contents

- Advanced Analytics At Roche
- RAAD Challenges – Increasing engagement in Advanced Analytics
- Crowd Sourcing Advanced Analytics

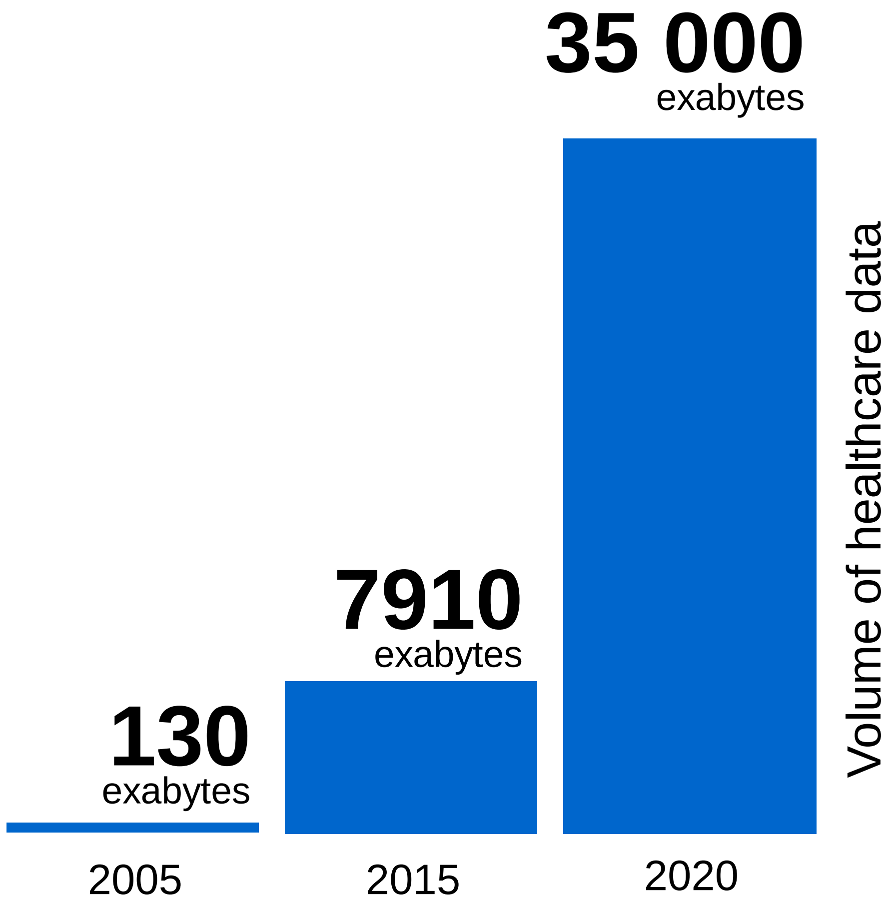
Acknowledgements

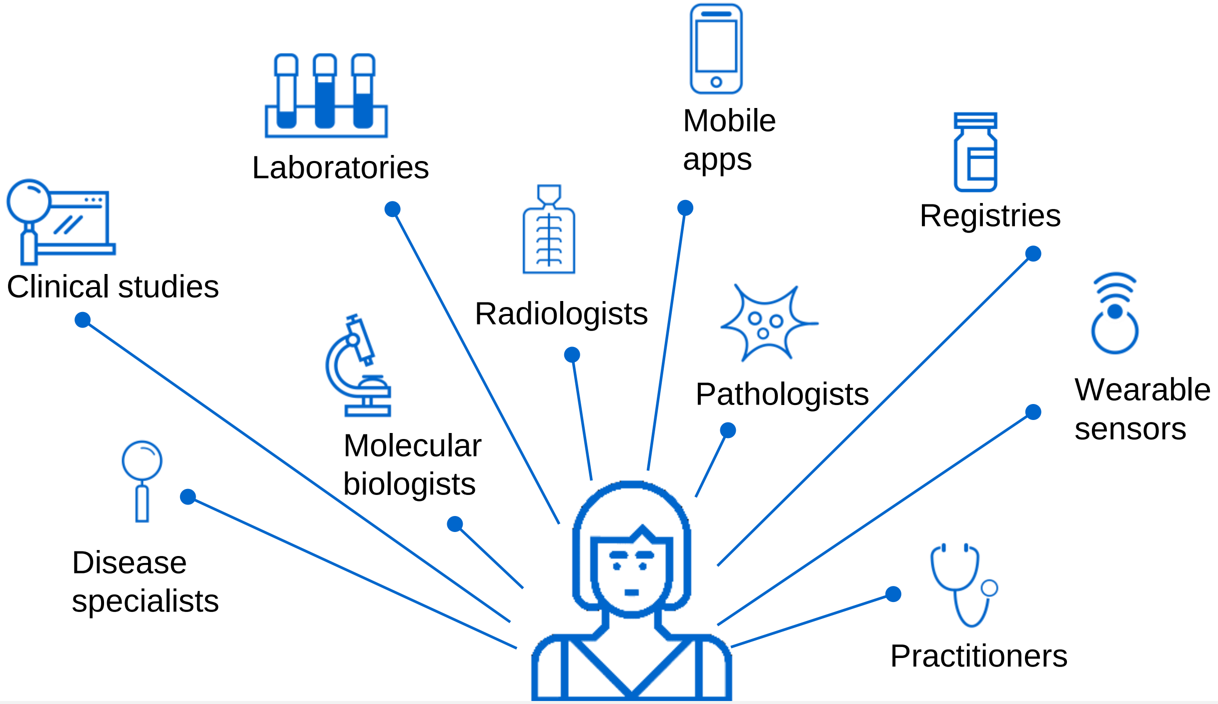
Roche

- Bharati Kumar (Intern)
- RAAD Challenge Organising Team & Participants
- James Black (RAADC Leader)
- Ryan Copping (RAAN Chair)

Technology is increasing the volume and the variety of healthcare data[1,2]

Linked, multi-modal and longitudinal patient-level data at scale provide the potential for a more complete view of patients, their journey and disease



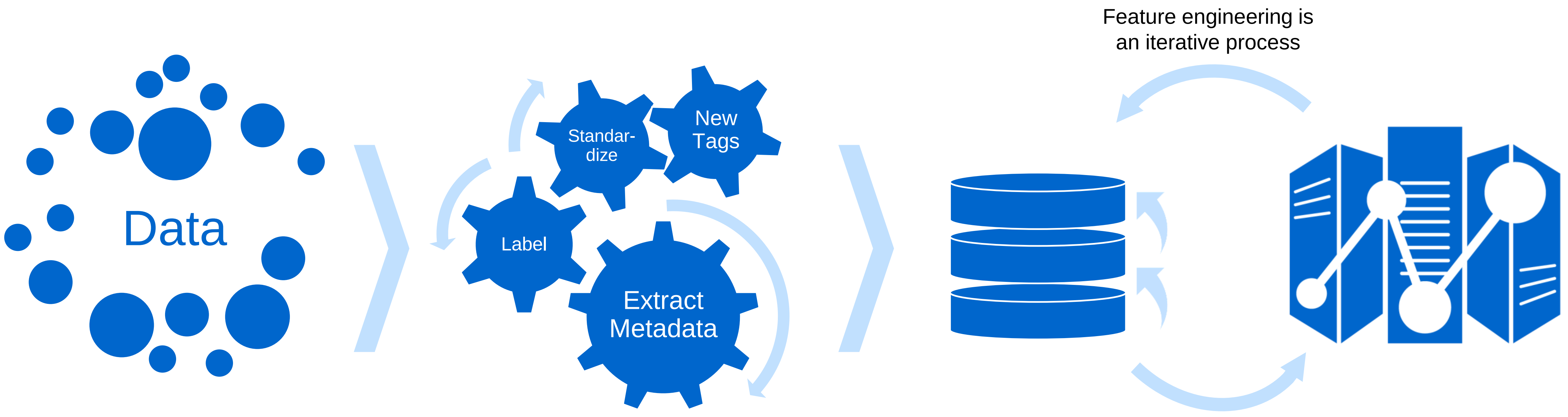


Vast numbers of data points from an increasing variety of sources now exist for individual patients in real world and clinical trial settings. Most real-world data are fragmented across providers, unstructured and separated over time – efforts to link and curate these datasets provide a huge opportunity to advance healthcare.[2]

1. Feldman B *et al.* 2012. Big data in healthcare hype and hope. Available from: https://drbonnie360.com/white-papers. (Accessed 8 August 2019);
2. Raghupathi W, Raghupathi V. *Health Inf Sci Syst* 2014;2:3

Making sense of complex healthcare data requires a considered approach

Significant time is invested in expert curation and preparation of the data before advanced analytics can be applied

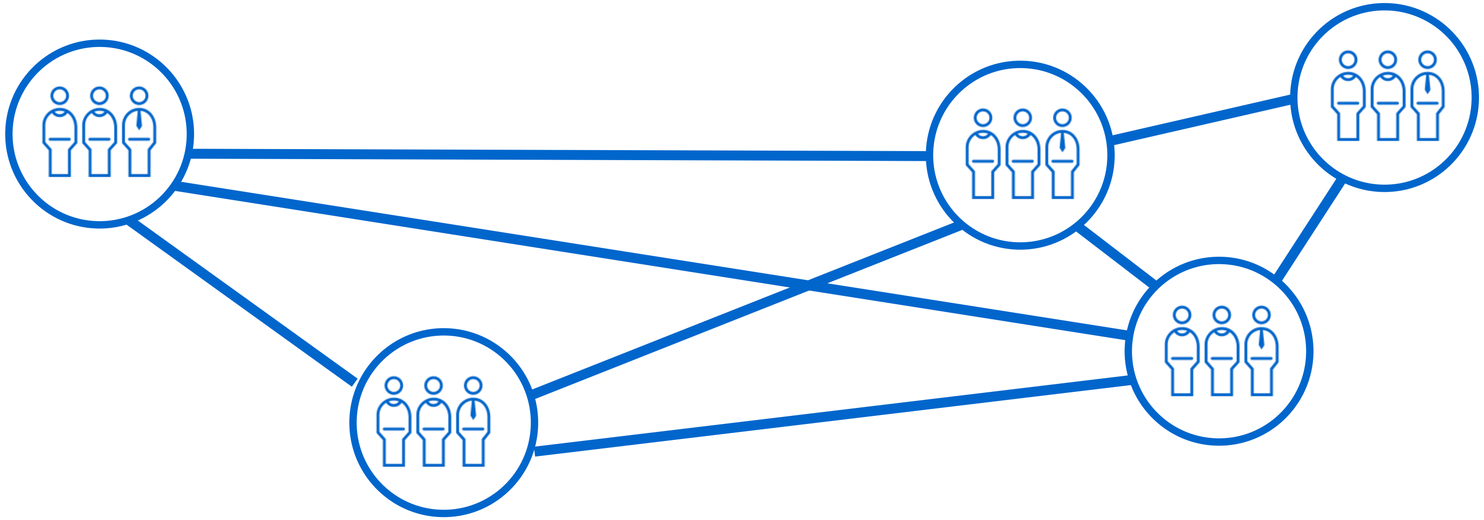


Acquisition	Curation	Feature engineering & Advanced analytics
High-quality data must be systematically collected[1,2]	Relevant data must be carefully categorized, linked and curated[1,2]	Data must be transformed based on understanding of the question, biology and analytic approach to create relevant features. Analytics must be carefully designed to address specific research questions

1. Kruse CS *et al. JMIR Med Inform* 2016;4:e38; 2. Anagnostopoulos I *et al. J Supercomput* 2016;72:1494–1516

Roche is integrating clinical trial data to create large, high-quality, disease-specific data marts

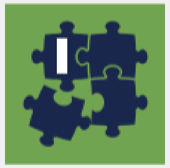
Roche



Specific data marts have been built for cancer immunotherapy (CIT), hematology, autism and asthma. The CIT data mart alone includes clinical, genomic and outcomes data from ~20 clinical trials

Findability
Where are our data?

Accessibility
How can I get the data?



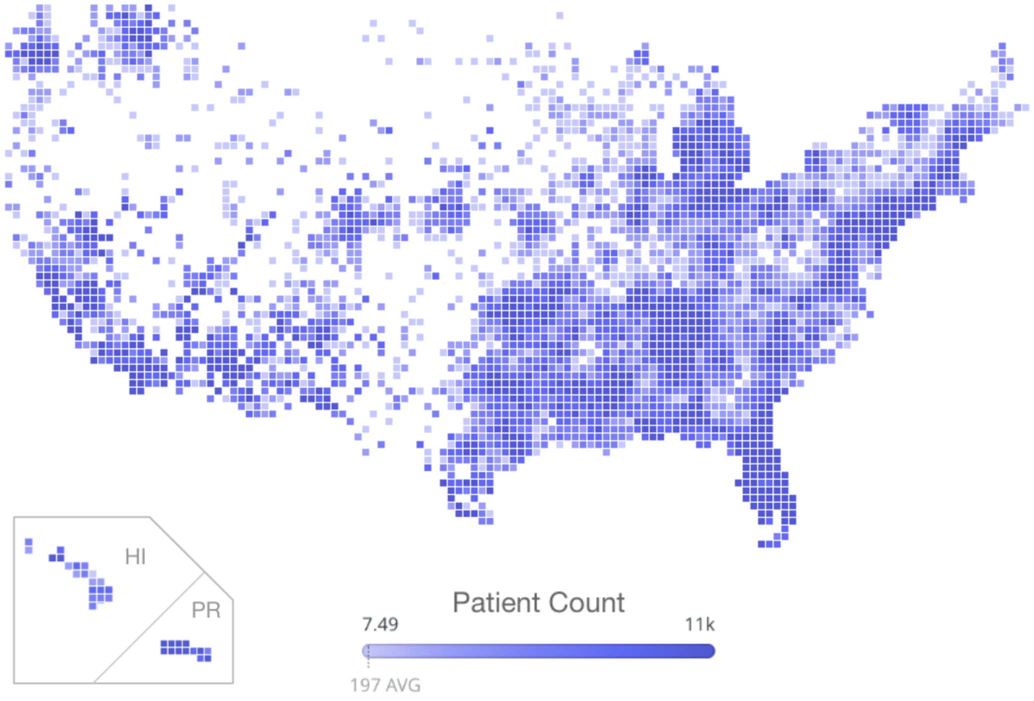
Interoperability
How can I connect or integrate the data?

Reusability
Can our data be easily shared and used again?

Our Enhanced Data and Insights Sharing (EDIS) programme integrates our vast repository of clinical trial data. With time, new data will be added prospectively, creating data marts of ever-increasing size

Roche is building a number of key partnerships to advance research grade RWD

Example: Flatiron Health



The Flatiron database provides an opportunity to gain insight into the 97%[1,a] of cancer patients in the USA who are typically not enrolled into clinical oncology studies

Curation of this diverse and rich database, and definition of relevant real-world endpoints, is the first step in the journey to maximize its potential in research and healthcare

Curation of Flatiron datasets will ensure research grade, RWD are made available across the industry to accelerate research and development, access and personalised care

2.2 M active patients

2500 clinicians

280 cancer centres

80 sites of care

flatiron

[a]Based on an internet survey of 2027 adults with cancer

RWD, real-world data

1. Comis RL *et al. J Oncol Pract* 2019;5:50–6

Roche is enabling linkages of high-quality, multi-modal data

Example: The Flatiron/FMI clinico-genomic database is an enriched, industry-leading research platform

Roche

flatiron

- Laboratory values
- e-Prescriptions
- Demographics
- Diagnosis
- Visits

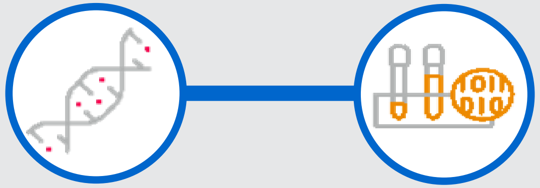
Structured data processing

- Discharge notes
- Physician notes
- Pathology
- Radiology reports

Unstructured data processing

- Overall survival
- Real-world PFS
- Real-world response

Clinical outcomes



The clinico-genomic database contains longitudinal data for more than 40 000 patients, and will continue to grow and expand overtime

FOUNDATION MEDICINE

Genomic data

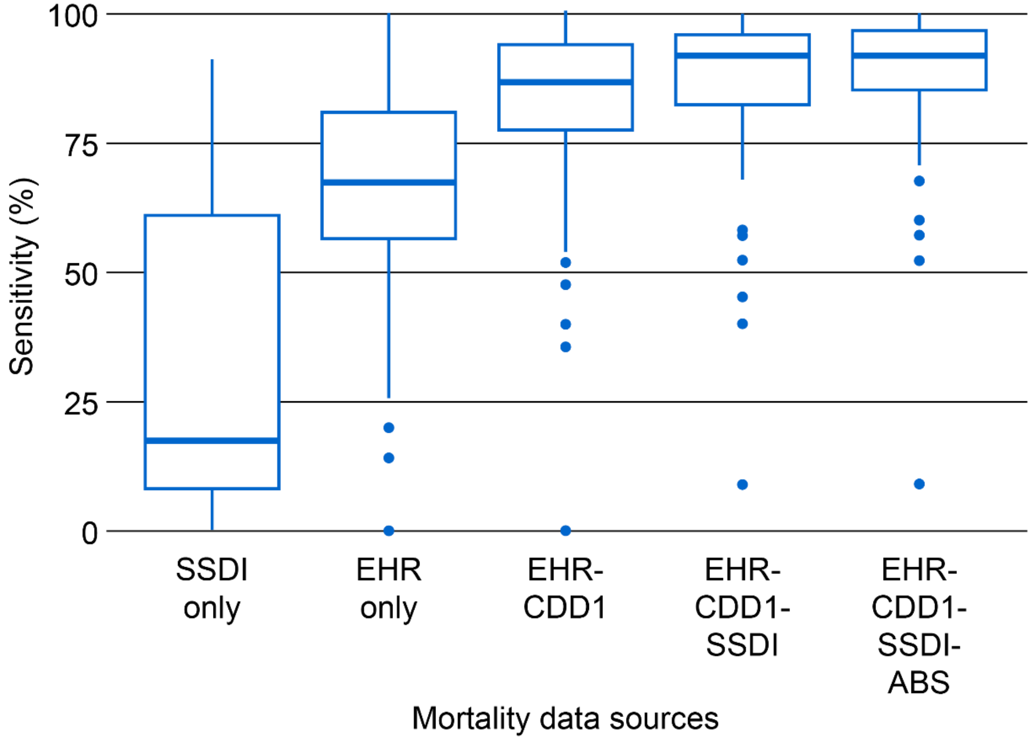
Foundation medicine is a leading molecular insights company specializing in high-quality, comprehensive genomic tumour profiling

FMI, Foundation Medicine Incorporated; QoL, quality of life

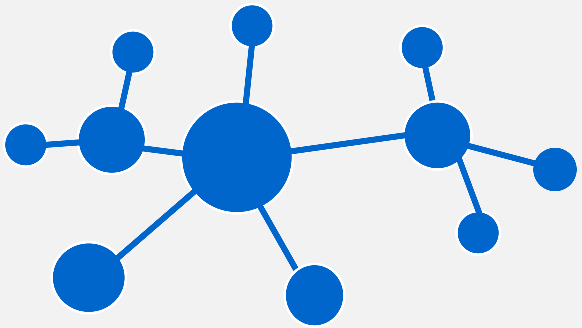
Improving data quality is pivotal for rigorous scientific research

Example: Sensitivity of Flatiron mortality data increased from 66% to 91% with additional sources[1]

Sensitivity of mortality data for advanced non-small-cell lung cancer



Flatiron curates and abstracts both structured and unstructured data to improve data quality[1]



Amalgamating data sources resolves data gaps and improves data reliability[1]

Box plots show median sensitivity. Upper and lower hinges correspond to 25th and 75th percentile. Whiskers represent sensitivity within 1.5 IQR of lower and upper quantiles, with points outside the whiskers showing residual data

ABS, abstracted; CDD1, commercial death data; EHR, electronic health records; IQR, interquartile range; NSCLC, non-small-cell lung cancer; SSDI, social security death index. 1. Curtis MD *et al. Health Serv Res* 2018;53:4460–76

Early Experience : Predicting oncology outcomes from clinical trial data

Project Turing: Evaluating advanced analytic approaches

The goal

Evaluate and compare approaches to predicting patient outcomes in DLBCL

Prediction of OS, PFS and treatment response using baseline clinical and genomic data

Compare the performance of externally-developed, proprietary machine learning models with Roche's open source-based approaches

The data

Cleaned de-identified clinical trial data on DLBCL from a Phase 3 trial that included > 1400 patients

DLBCL, diffuse large B-cell lymphoma; OS, overall survival; PFS, progression-free survival

Early Experience : Project Turing

Results: Simple, interpretable, machine learning models may may be more appropriate than more complex models depending on the scale of the data

The results

Outcome	Performance metric	Single biological feature	Open source ML models with internal biological knowledge	Proprietary machine learning models	
OS	Weighted C-index	0.56	**0.66**	0.62	0.60
PFS	Weighted C-index	0.64	**0.67**	0.45	0.64
ORR	AUC	0.68	**0.70**	0.68	0.55

Key Learnings

Combining machine learning and biological understanding generated the best predictions

Predictions and accuracy were relatively consistent for all approaches

Many expected variables appeared, plus some novel ones

Similarity in outcomes is likely because of the size, depth and completeness of the data

AUC, area under the curve; ML, machine learning; OS, overall survival; PFS, progression free survival; ORR, overall relative risk

Roche Advanced Analytics Data Challenge 1.0

The goal

Predict the probability a patient will be alive at 1 year after treatment initiation, using all the patient data available up to the start of treatment

The why

Demonstrate the potential of applying advanced analytics to RWD to strengthen R&D decision making

Bring Roche talent together on an analytic problem

The data

7000 flatiron patients electronic health records were used across seven cancer types to build a model

Average patient has **~190** data points pre treatment

3500 fresh patients used to test the models

500 Roche employees

132 teams

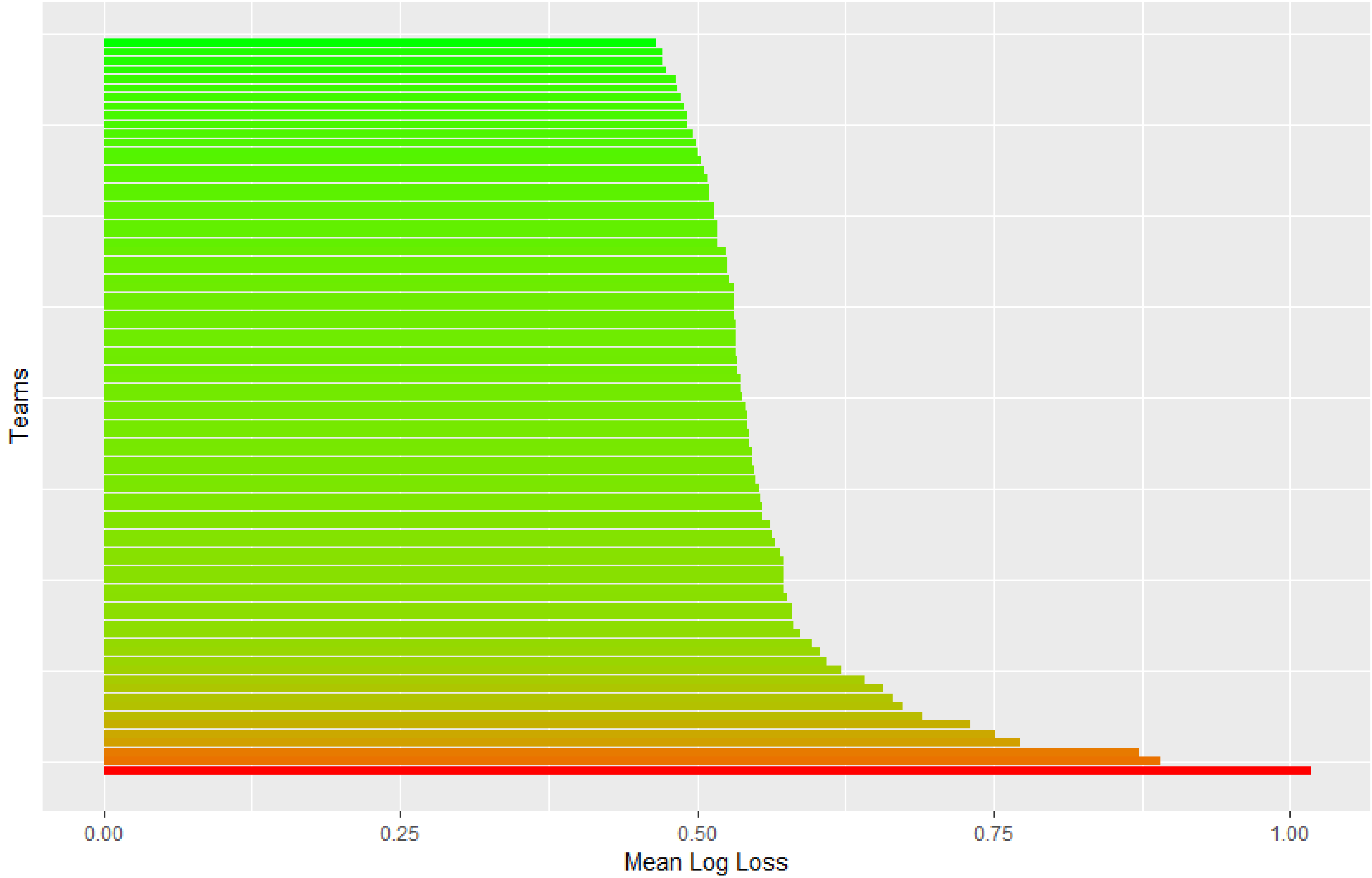
165 departments

28 Roche sites

61% of teams submitted a model

MDAS, meaningful data at scale; RWD, Real World Data

RAAD Challenge 1.0 Results

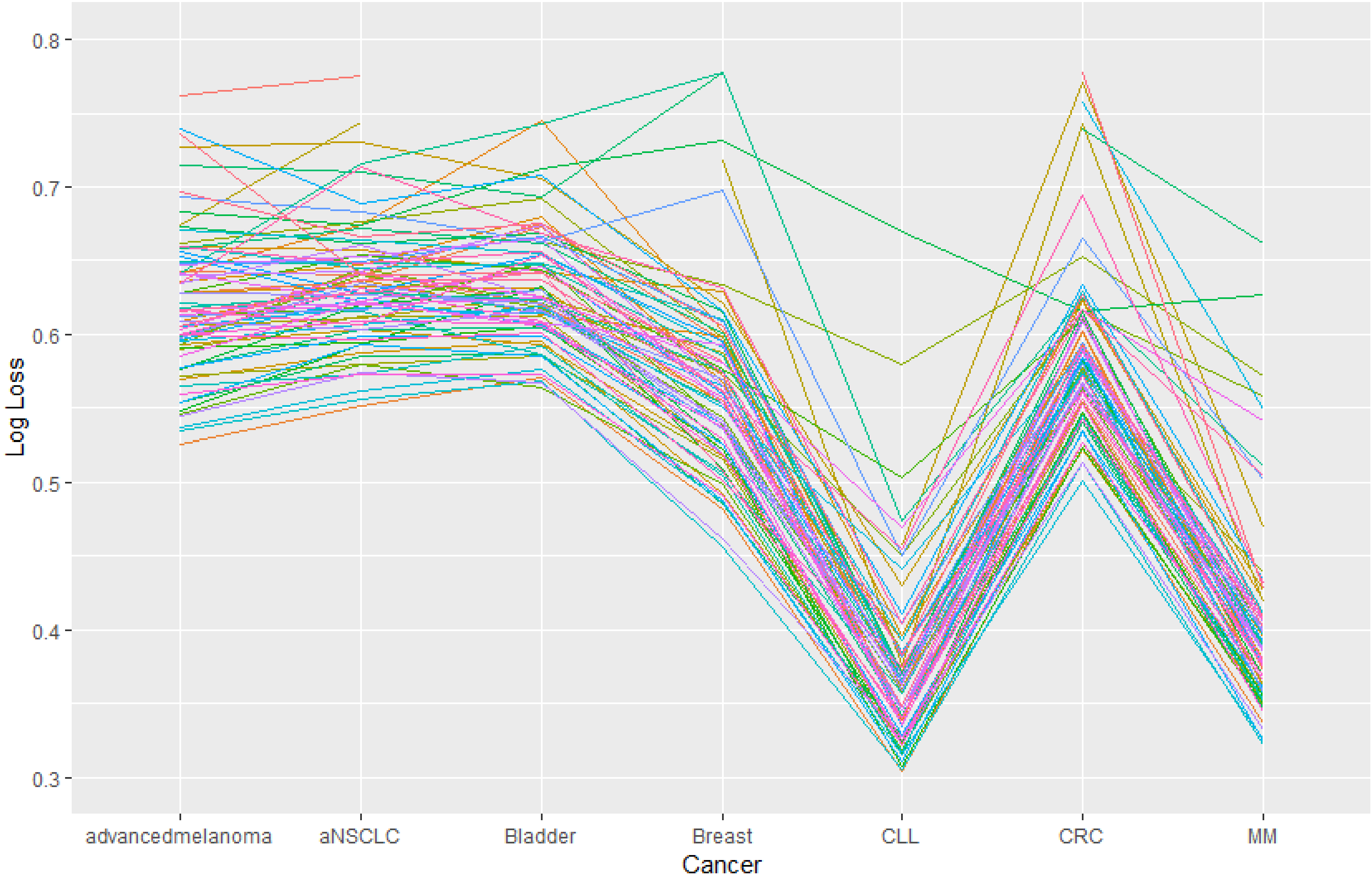


The approach

Most of the winning teams used XGBoost, a non-parametric tree-based algorithm

There was wide variety in the way models were applied and pre-processing was performed

RAAD Challenge 1.0 Results



The approach

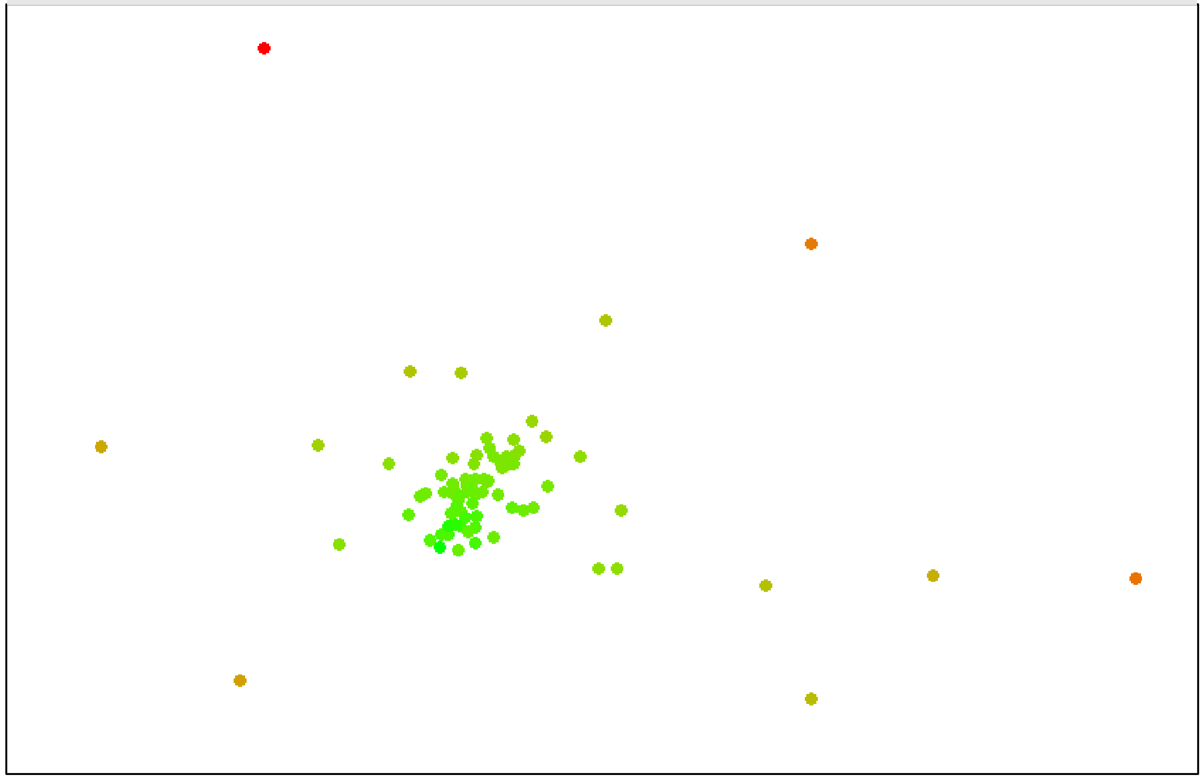
Most teams fitted a common model across all cancer types

The accuracy of predictions from different teams was similar across different cancer indications

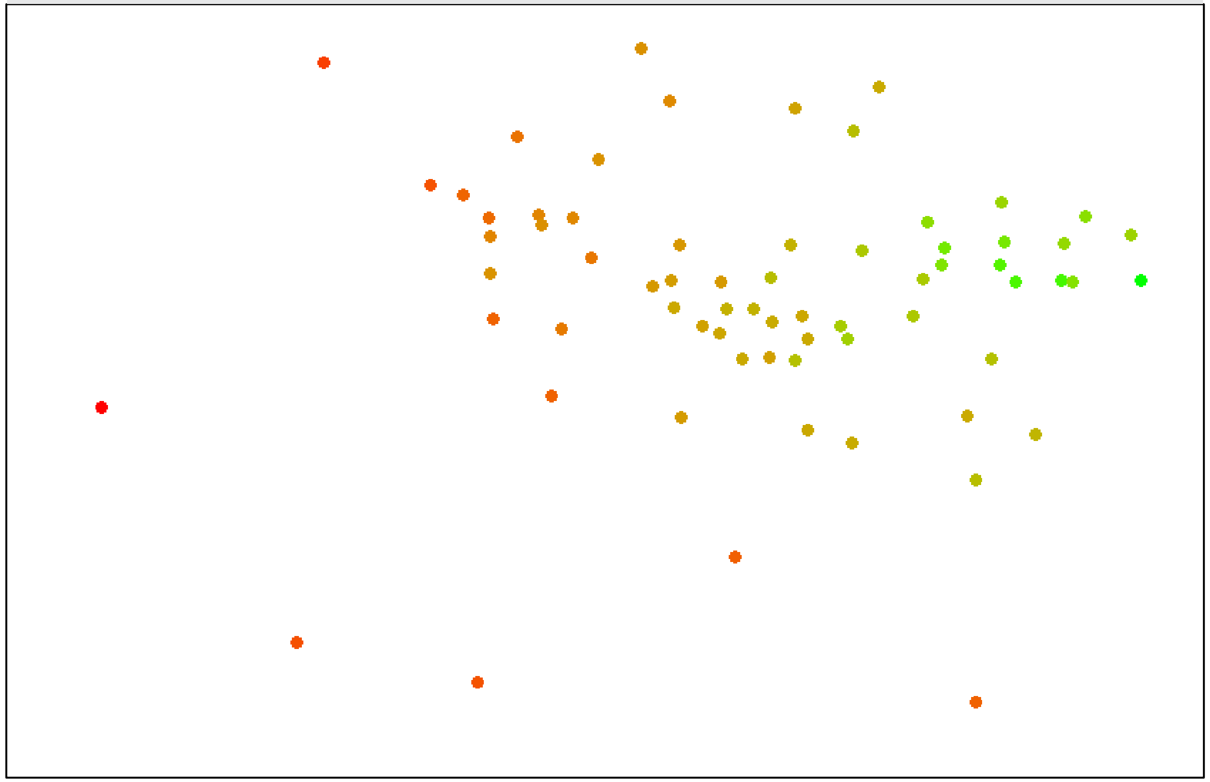
RAAD Challenge 1.0 Results

Multidimensional Scaling of Teams Based Upon Their Predictions

All Teams

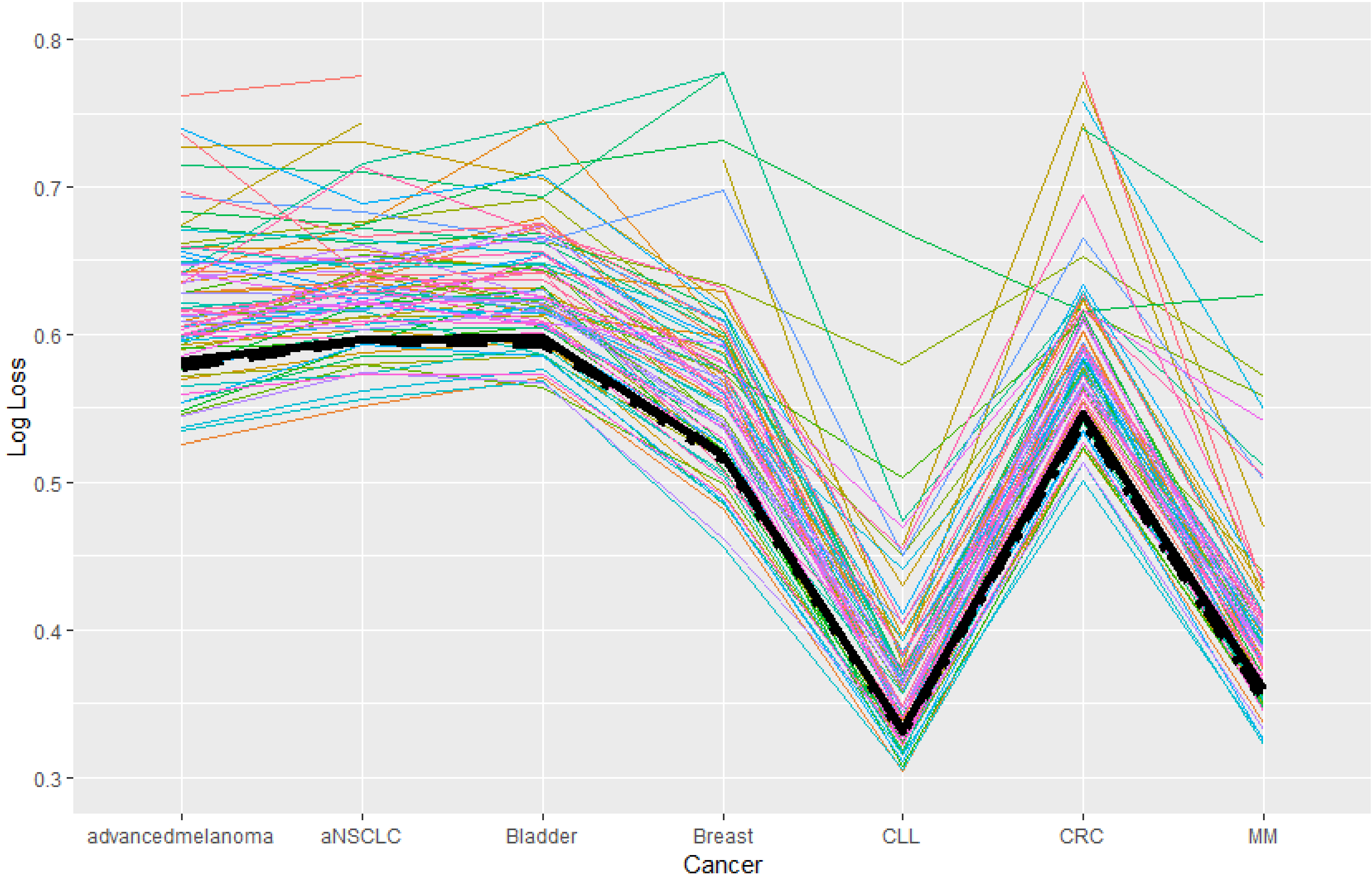


Better Performing Teams



The further your predictions are away from the winning team, the worse your score

RAAD Challenge 1.0 Results
Consensus Scoring



Consensus Scoring

Taking an average prediction probability over all teams for each patient does not beat the best teams but gives a better than average score

This can be marginally improved by taking a more robust median or trimmed mean

Roche Advanced Analytics Data Challenge 2.0

The goal
To use advanced analytics to develop a prediction model to identify patients with NSCLC at first line most likely to respond to Tecentriq® treatment vs SOC

The data
- Training data set: 10 EDIS curated clinical trials (~5000 patients)
- Test data set: 1 clinical trial with minimal indication overlap (~1000 patients)
- Clinical and omics data provided


Enhanced Data & Insights Sharing (EDIS)

Timeline : 2 Months

517 participants

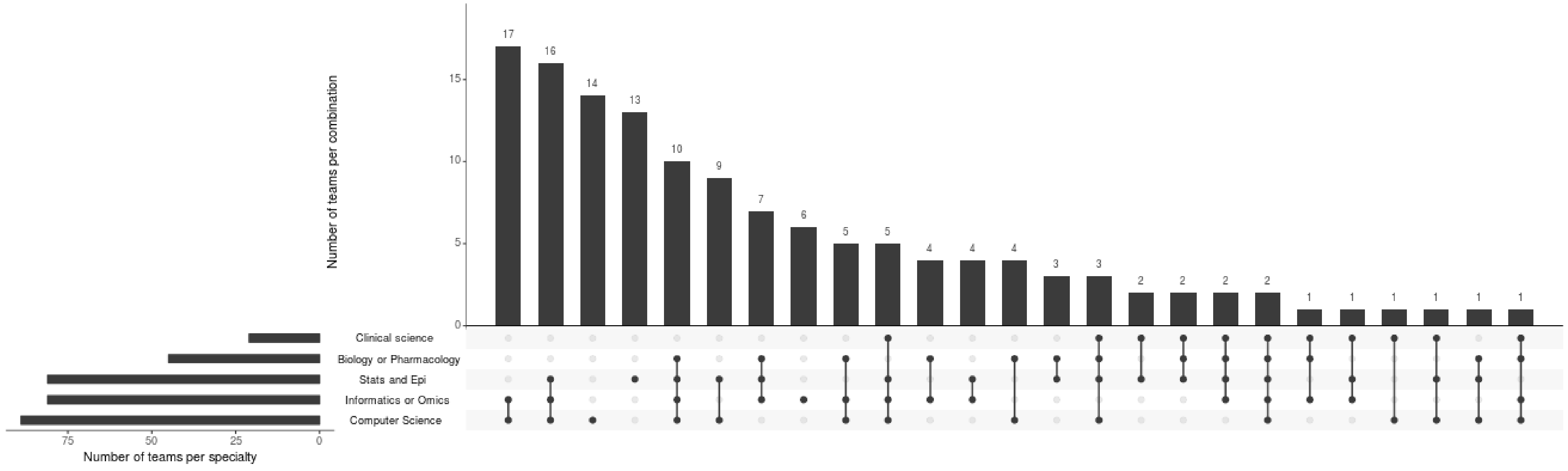
141 teams

38 Roche sites

85 teams submitted a final model

NSCLC, non-small-cell lung cancer; SOC, standard of care

Roche Advanced Analytics Data Challenge 2.0
Successful in Promoting Cross-Disciplinary Interactions



Roche Advanced Analytics Data Challenges Success Features

High Profile
Prize of lunch with CEO

Competition
Leaderboard on subset of held-out data

Community
Discussion Forum
Teams had to be cross-functional – "Dating service" to promote this

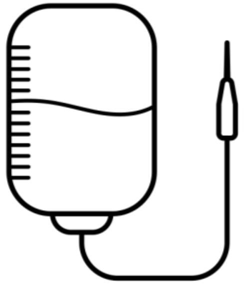
Accessible
Self-contained problem
Available easy-to-access data
Available compute environments

Focussed
Restricted timeline of 2 months

Multiple Awards
Model Performance
Best Advanced Analytics Practice
Scientific Insights

RAAD Challenge 2.0
Data & Scoring Metric

VS



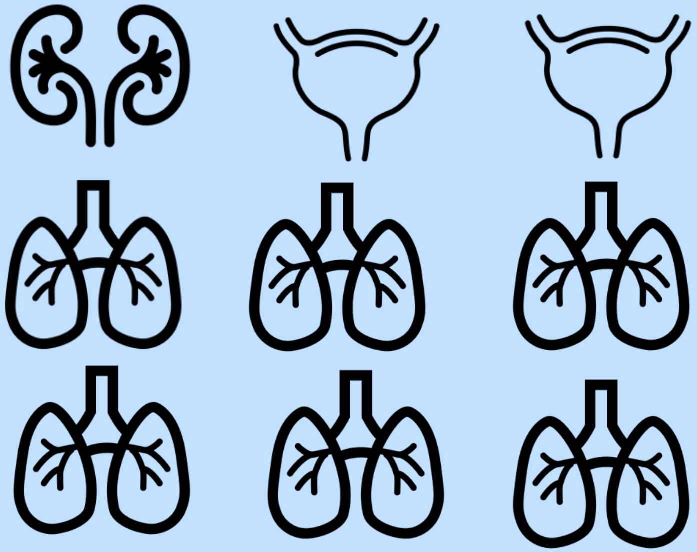
Prevalence of selected subset in range 20%-80%

Roche

Predictions From Teams	Blinded From Teams

Training Data

9 Studies in Various Indications



Test Data

1 Study in Lung

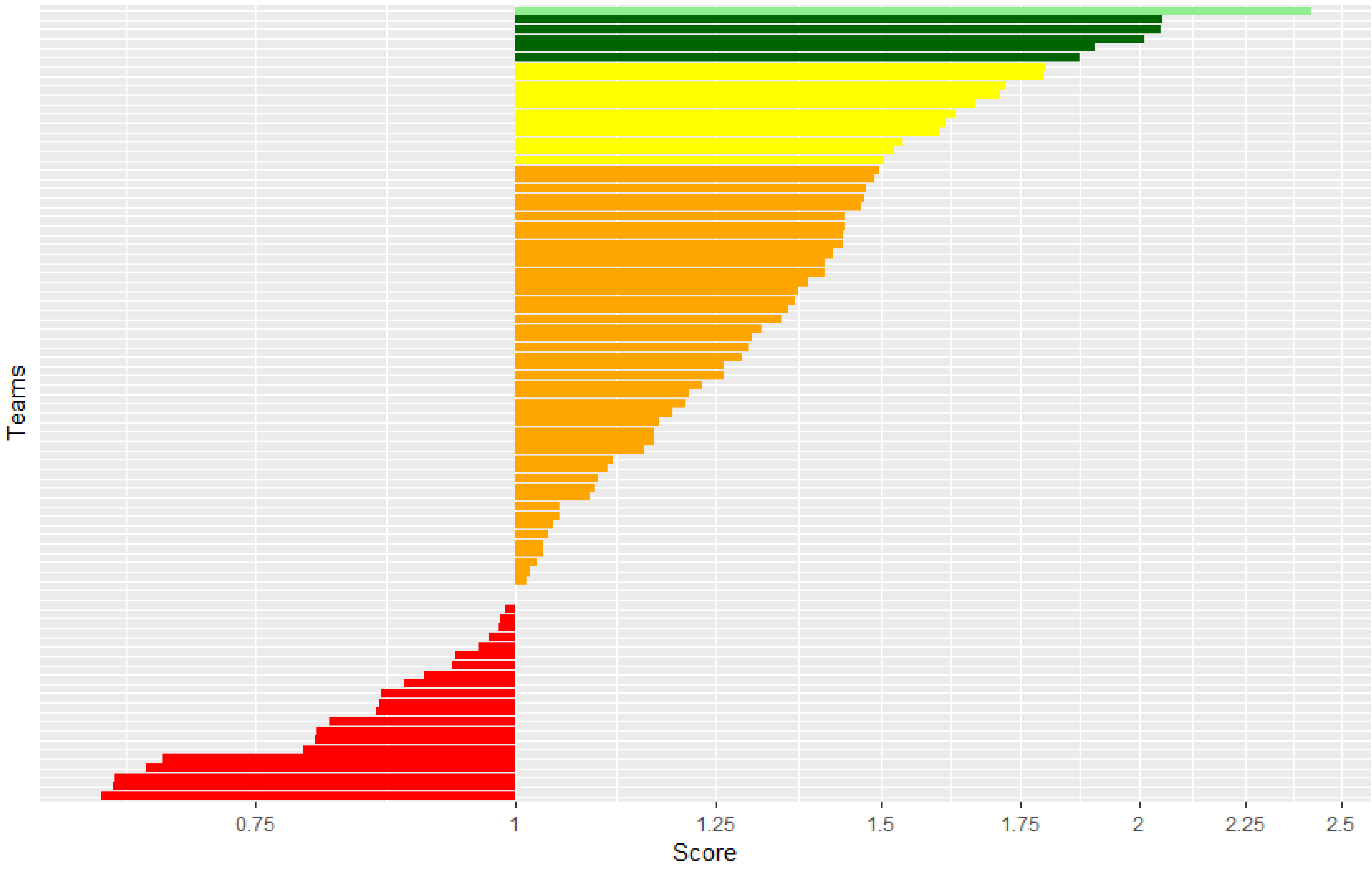
Randomised immunotherapy vs Chemotherapy

Patient	Subgroup	Treatment	Response
001	Y	CIT	Y
002	Y	CIT	Y
003	N	CIT	N
…..	…..	…..	…..
998	N	Chemo	N
999	Y	Chemo	N
1000	Y	CIT	Y

$$\frac{\text{Odds Ratio}_{\text{Selected}}\ \text{Tecentriq vs Chemo}}{\text{Odds Ratio}_{\text{Unselected}}\ \text{Tecentriq vs Chemo}}$$

```
glm(Response ~ Treatment + Subgroup + Treatment*Subgroup , family=binomial)
```

RAAD Challenge 2.0 Results



The results

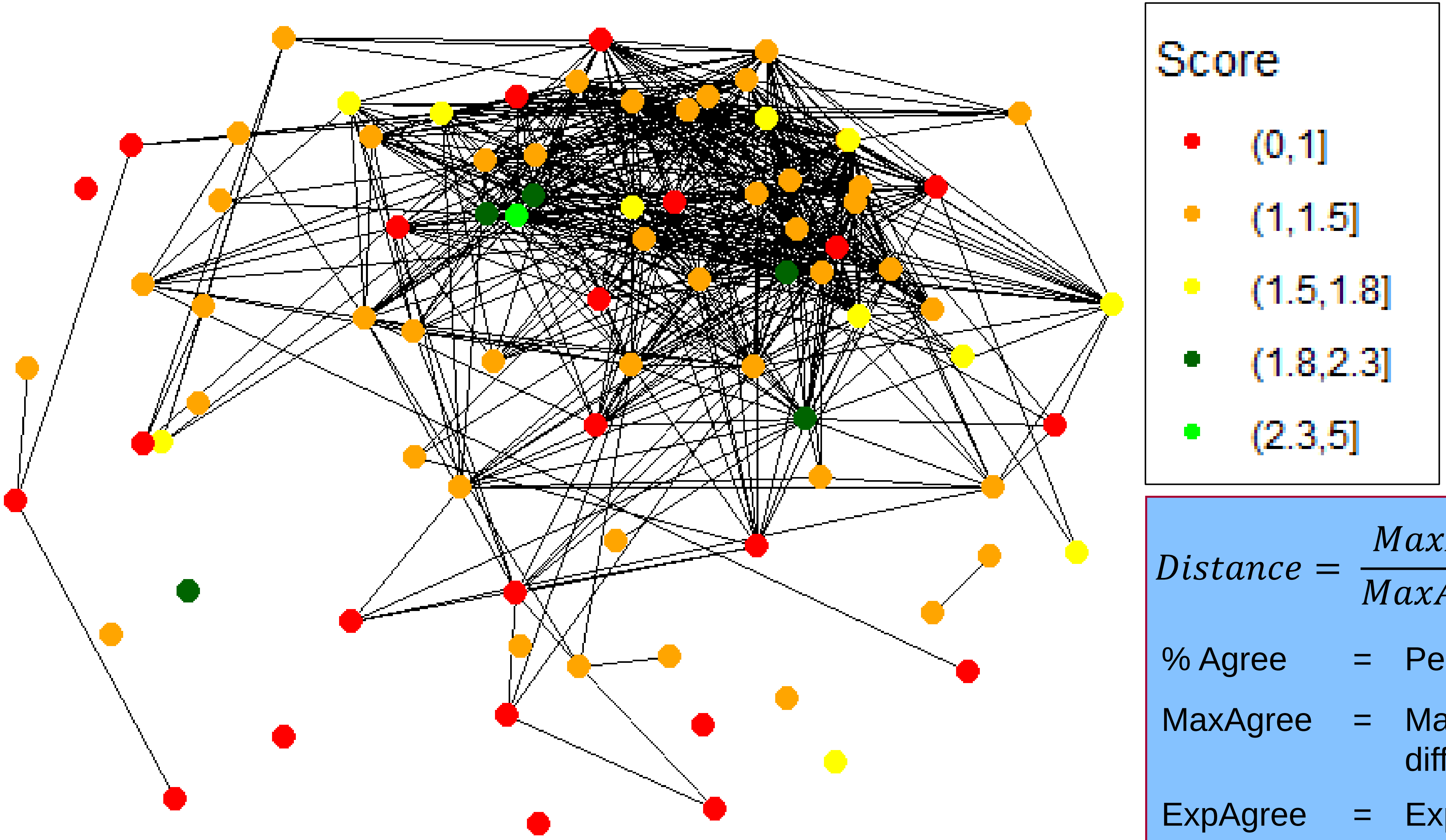
Teams deployed a wide variety of different methods and approaches

The overall winner was a simple model reflecting the common biology across indications

When predicting treatment responses across a variety of different studies and indications, model robustness is critical

RAAD Challenge 2.0 Results

Multi-Dimensional Scaling of teams from their predictions, allowing for differences in sizes of selected population



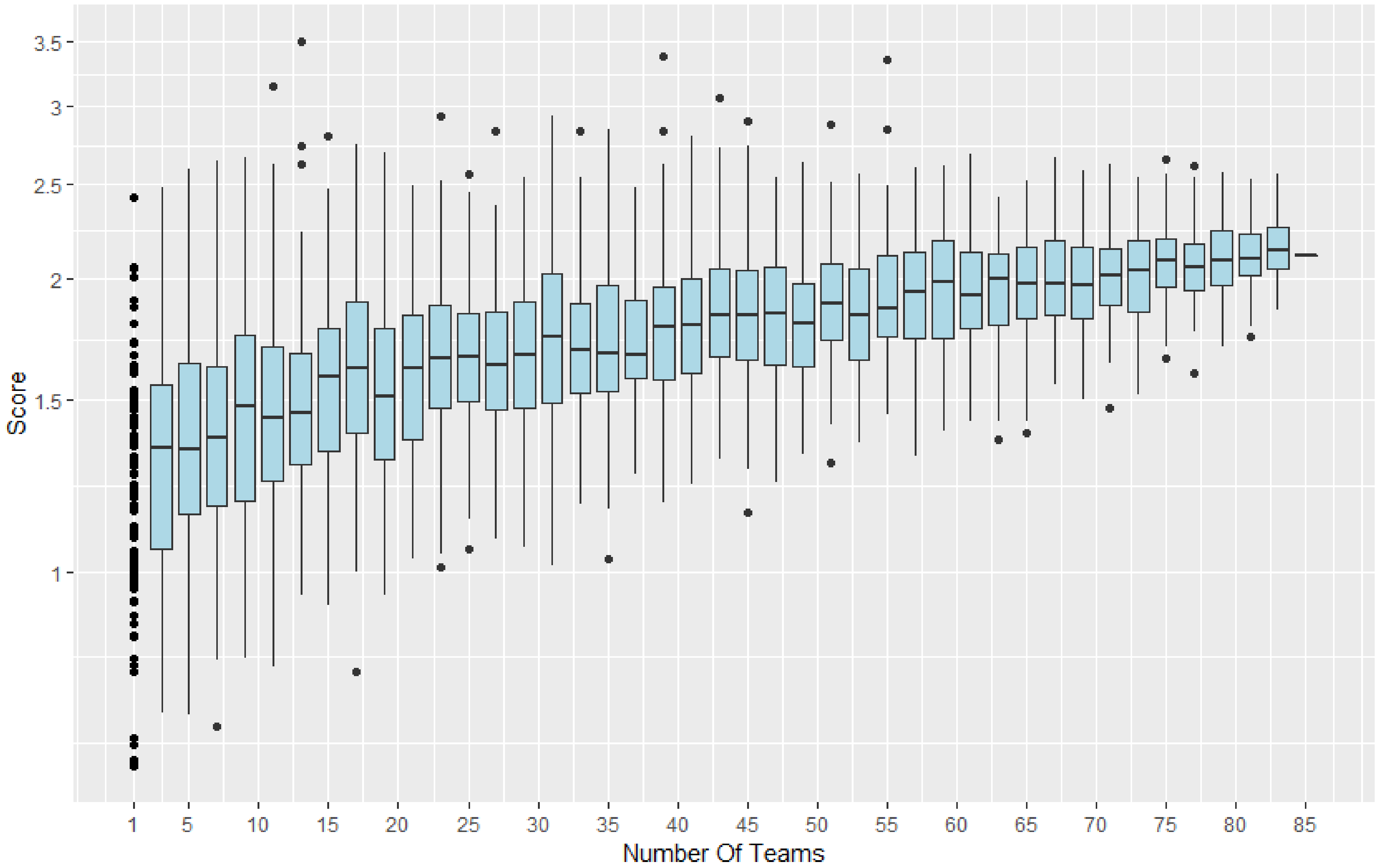
$$Distance = \frac{MaxAgree - \%\,Agree}{MaxAgree - ExpAgree}$$

% Agree = Percentage Agreement

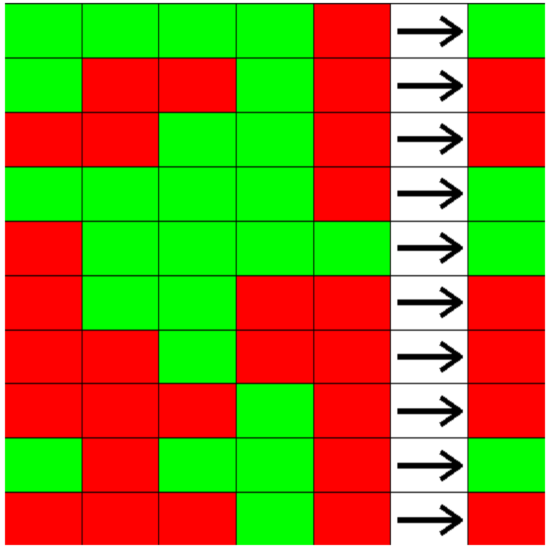
MaxAgree = Maximum agreement given difference in prevalences

ExpAgree = Expected agreement by chance

RAAD Challenge 2.0 Results
Consensus Scoring Of Predictions

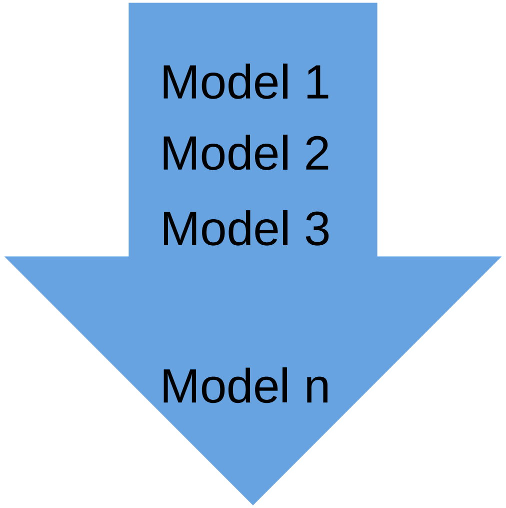


- **Randomly sample n teams**
- **Determine selected population by simple majority vote across teams for each patient**



Why Might Consensus Scoring Work?
Two Related Ideas

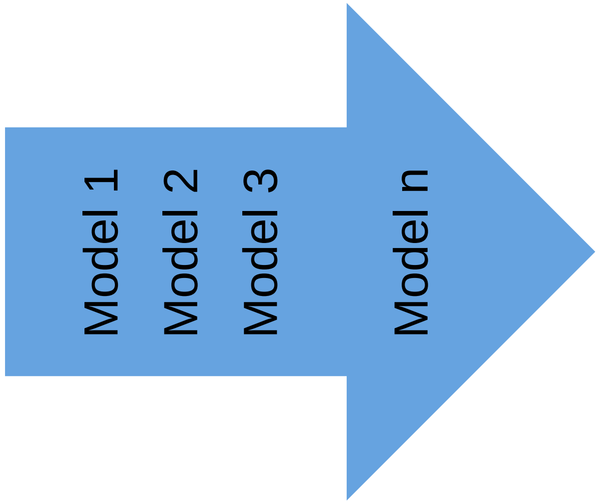
Bagging
e.g. as used as a component of Random Forests



Aggregated Model

The average of a set of unbiased high variance models will be unbiased with lower variance

Stacking

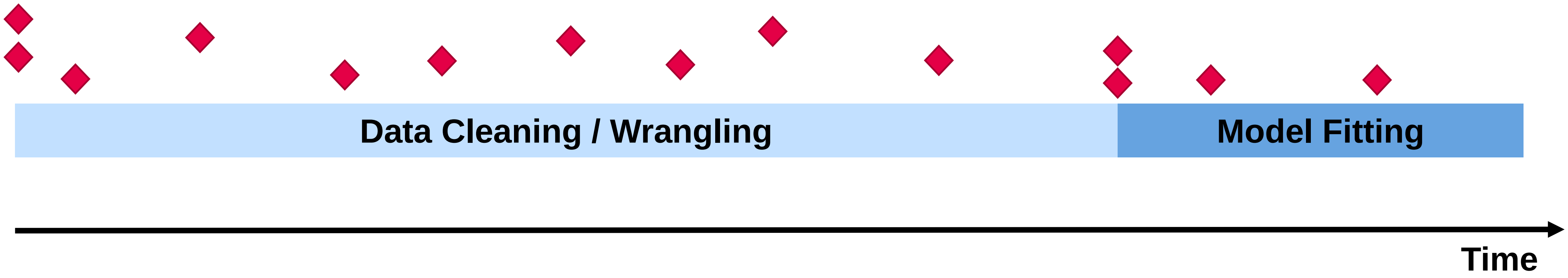


Meta Model

Regressing on the predictions from individual model gives a meta-model with frequently better performance

Why Might Consensus Scoring Work?

We make lots of decisions throughout the whole modelling process
Most decisions are made in data cleaning before starting modelling

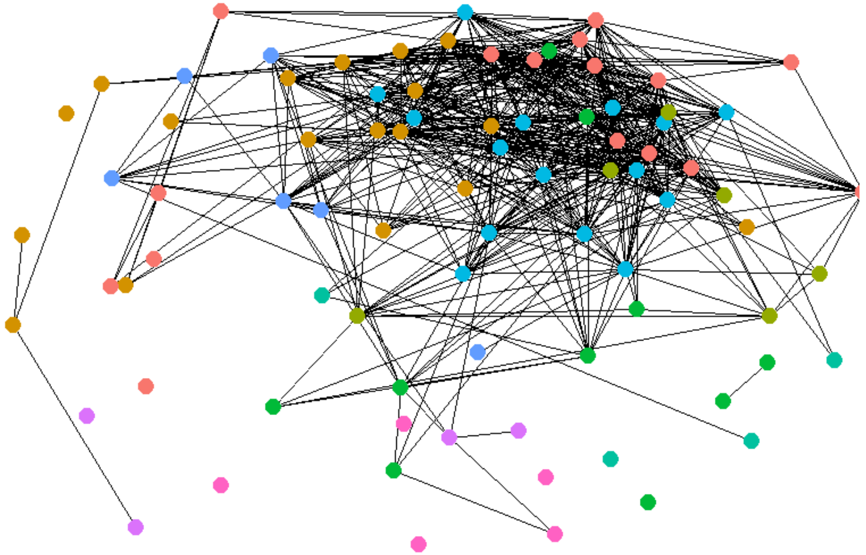


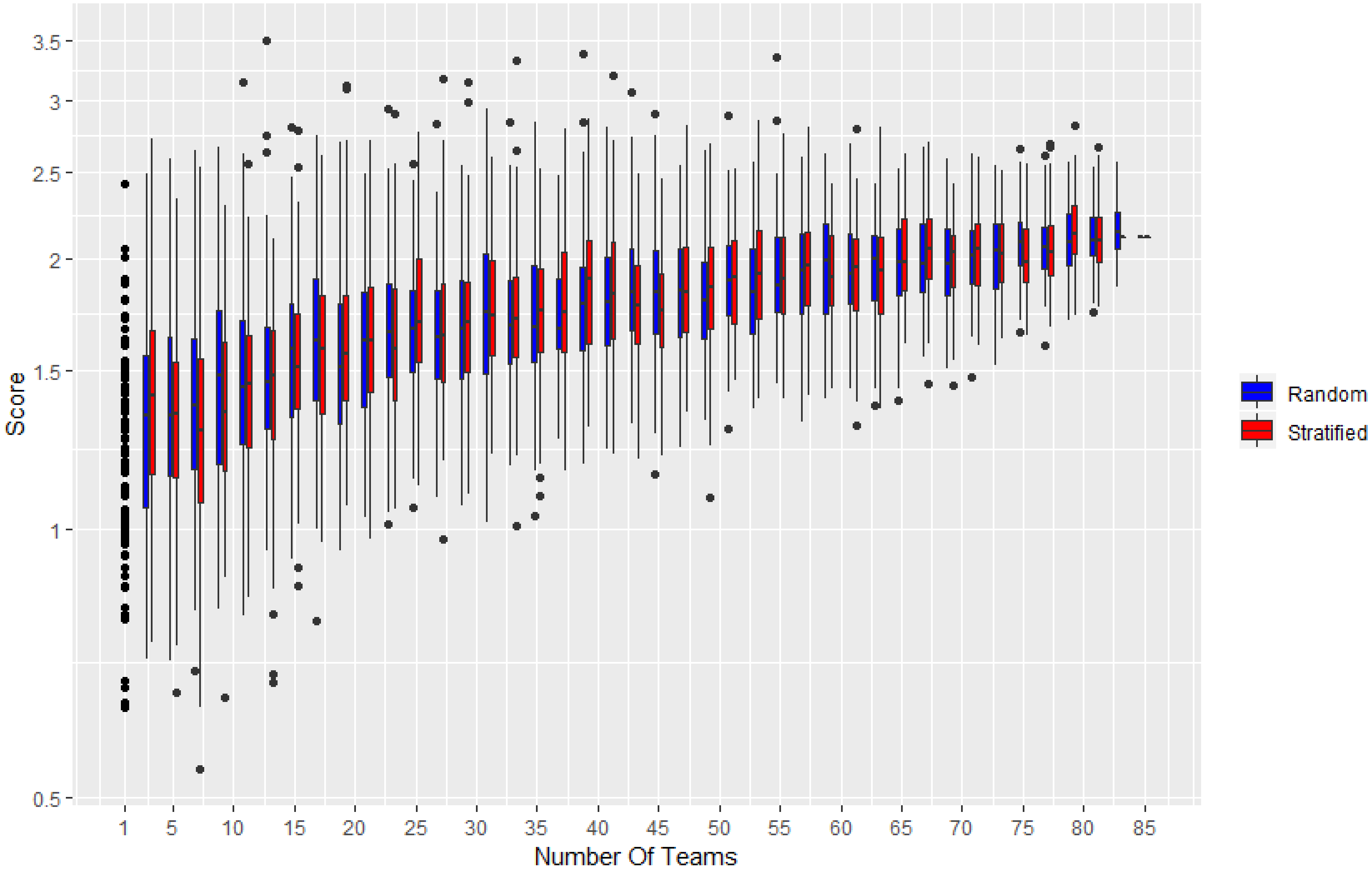
"How do I impute missing values?"
"Which variables do I delete through too many missing values?"
"Do I treat as categorical or continuous?"
"Which time-points do I use?"

"Which technique do I use?"
"Which parameters do I optimise?"

Most questions don't have a right/wrong answer
All questions will impact the fitted model, generally in unpredictable ways
Considering learners covering the whole of the modelling process increases variability, beyond what you obtain by just varying the modelling process potentially giving us better predictions.

RAAD Challenge 2.0 Results
Consensus Scoring Of Predictions, Data Stratified Sampling





- Complete linkage clustering of teams into 9 clusters
- Stratify random sampling of teams by cluster

Other ways of stratifying teams, e.g. by methods or data used, or selecting teams are being investigated

Final reflections

Investment In Data and Analytics

- Data Sources
- Data Curation & Integration
- Data Analytics

Data Challenges Have Been Successful Beyond Expectations

- High Profile Promoting Advanced Analytics Across the Company
- Development Opportunity for Individuals
- Shared Learning Activity for Teams
- Promoting Cross-Disciplinary and Cross-Organisation Interactions

Consensus Scoring Has Potential Within Modelling

- Further Investigation Required
- Can Concept be Automated to Make it Feasible Within Resource Limitations


Roche
Thank You For Your Attention
Questions?