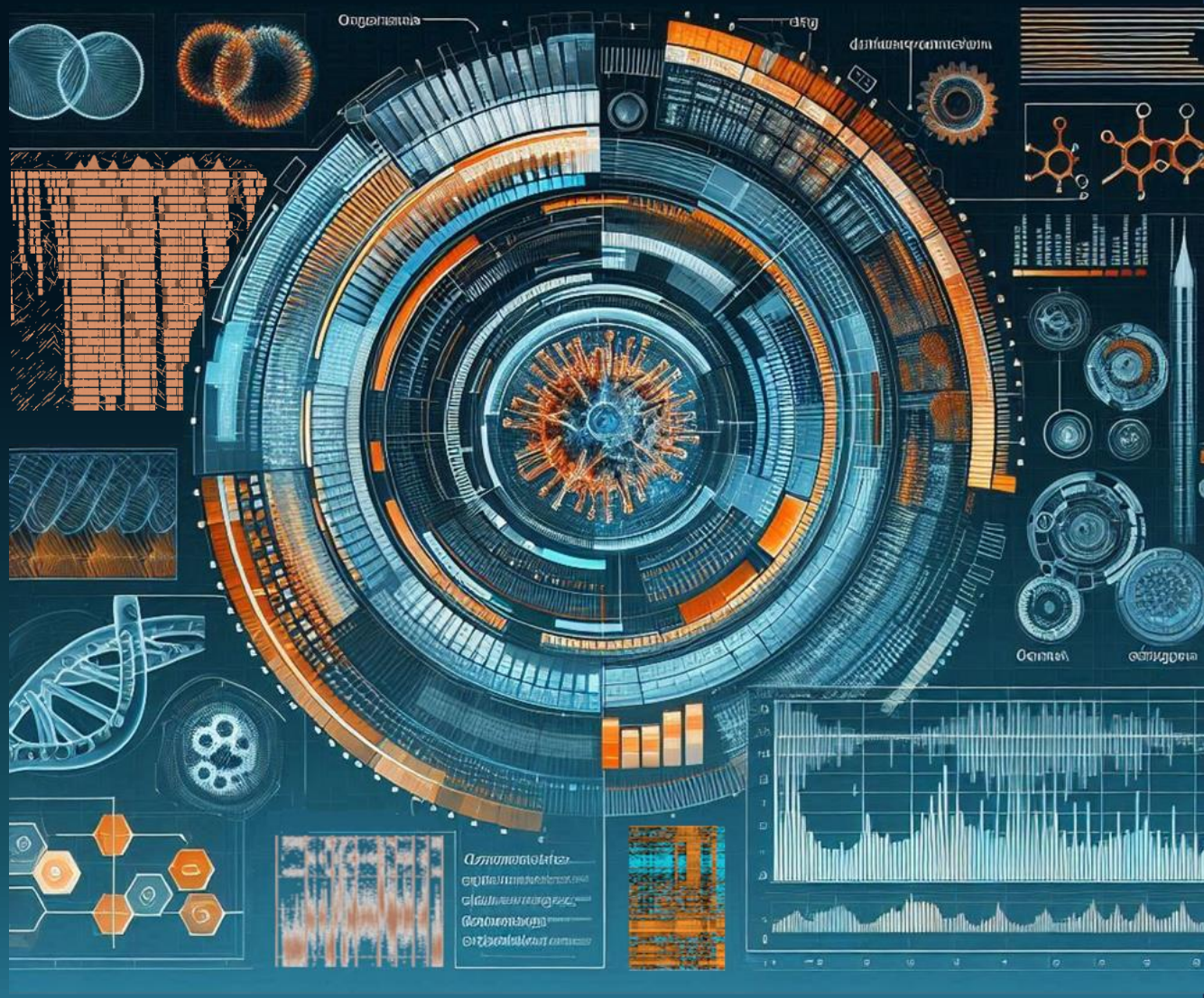
MAQC Society

Machine Learning Working Group Kick-off Meeting

Wendell Jones, PhD

Principal Bioinformaticist and Scientific Advisor
IQVIA Laboratories Genomics

MAQC Society Executive Chair
https://themaqc.org



Welcome to the MAQC Society Machine Learning Working Group

Introductions

Due to time constraints, after I call on you, please give your

- Name
- Title or Position or Student status
- Institution

MAQC Society

Agenda

1. Introduction, background, and plans for the working group
2. Goals and objectives
3. Descriptions of the first data to be used in the study
    a. Breast Cancer tabular dataset – OMOP CDM-like
    b. Lung Cancer tabular dataset - OMOP CDM-like
4. Requirements of participants
    a. Signatures on non-disclosure forms
    b. Fill out Data Analysis Plan template (2 pages) and submit to Steering Committee
    c. Train ML and/or stat models over 4 iterations of data disclosure
    d. Disclose new classification results for review in a timely fashion
    e. Make all code and models available for review and reproducibility/etc analysis
5. Options
    a. Participants can be individuals or teams
    b. Participants can work with both or only 1 dataset
    c. Participants can provide classifications using more than 1 ML model (ie, feel free to experiment, but we would ask you to identify what you feel is your best model)
6. Key decisions now
    a. Work with both datasets in parallel or serially (due to their great similarity of structure, recommend serially)
    b. Timetable – Suggest 2-3 week interval between training/classification iterations (3 weeks for first iteration, then every two weeks after that)
    c. Decide on thresholds for survival for each dataset (trade-off is balance in classes vs. total number of patients in study)
    d. Decide on number patients which will have their outcomes discloses in each training iteration
        i. Recommend a schedule of roughly 100-150 for Iteration 1, 300-500 for Iteration 2 (mimic Phase 1-2 and Phase 3 clinical trial sizes), 50% in Iteration 3, 80% in iteration 4.
    e. Method of code archival
7. Key decisions later
    a. Examine other types of datasets (eg, imaging or more extensive molecular sets) before article-writing or construct article with what we have now for tabular data (suggested best practices)
    b. Number and variety of articles
    c. Measures of model reproducibility, performance, complexity, traceability, etc
8. Access to data and submissions

Background

- The MAQC Society history goes back to the MAQC Consortium and first publications on the reproducibility of microarrays which measured individual RNA expression in cells (2005-2006)
- The second phase of the MAQC (MAQC II) examined development of complex biomarkers from microarray data to predict outcomes (usually survival or toxicity) in multiple datasets using training/test methods. Many participants used ML methods (2010 *Nat Biotech, Pharmacogenomics*)
- With completion of several studies related to NGS measurement, the society now wants to examine ML methods as applied to human health, whether preclinical or clinical
  - How can we improve methods and procedures, architectures, etc. for ML models using human health data to predict outcomes or classify patients to elevate the process to "clinical grade" status?
- Our first dataset(s) in Phase 1 will involve real-world evidence (observational) data from EMR and other sources (anonymized) – Lung Cancer and Breast Cancer
- The Society has formed a Steering Committee for the ML Working Group and additional members will be added through different phases
  - Phase I - Dr. Wendell Jones (Executive Chair, MAQC Society and IQVIA Labs) and Dr. Russ Wolfinger (MAQC Board member, JMP Genomics) will be co-chairs
  - Phase I - Dr. Rebecca Kusko (Deputy Chair, MAQC Society and Cellino Bio), Dr. Joshua Xu (Executive Secretary, MAQC Society and US FDA), Dr. Samir Lababidi (US FDA), Dr. Benjamin Haibe-Kains (U. of Toronto) will be members

2025 MAQC Society Board of Directors

Co-chairs, Steering Committee

Steering Committee Members

MAQC Society

MAQC Society Board of Directors, Officers, and Senior Leadership

A Board of Directors is chartered to develop the Society and oversee the Society's activities. Specifically, the Board is dedicated to building and strengthening the reproducible science practice by (1) engaging scientists in academia, industry, and government, (2) fostering collaboration and harmonization of education, research tools, and new technologies, and (3) building the global-scientific infrastructure needed to advance and promote reproducible science.

Wendell Jones, PhD, Executive Chair

IQVIA Laboratories (Genomics), USA

Andreas Scherer, PhD, Past-President

Research Coordinator, Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland

Wenjun Bao, PhD, Treasurer

SAS Institute, USA

Joshua Xu, PhD, Executive Secretary

US FDA, USA

Ben Haibe-Kains, PhD, Board Member

Princess Margaret Cancer Centre, Canada

Rebecca Kusko, PhD, Deputy Chair

Cellino Bio, USA

Samir Lababidi, Board Member

Bioinformatician at FDA Office of the Commissioner

Russ Wolfinger, PhD
Director of Scientific Discovery and Genomics
Distinguished Research Fellow
Kaggle Grandmaster
JMP and SAS Institute

Xiaohui Fan, PhD, Past-President

Zhejiang University, China

Shraddha Thakkar, PhD, President

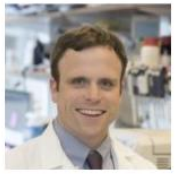
US FDA, USA

Leming Shi, PhD, Distinguished Scientist of the MAQC Society, Board Member, Chief Science Officer

Fudan University, Shanghai, China

Barbara Cheifet, PhD, Board Member

Nature Biotechnology



Chris Mason, PhD, Board Member

Weill Cornell Medicine, USA

Weida Tong, PhD, Chair Emeritus

US FDA, USA

Cesare Furlanello, PhD, Board Member

HK3 Lab S.R.L, Orobix Life & LIGHT CENTER BRESCIA, Italy

Aedín C. Culhane, Ph.D., Board Member

Professor of Cancer Genomics, University of Limerick

PANY CONFIDENTIAL

Overview of ML Working Group and Purpose

- Focus - Understanding properties and determining best clinical practice of ML models related to model evolution, reproducibility, explainability, traceability, complexity (data and model), and performance in relation to each other and in relation to
  - Dataset size (both number of observations and possibly covariates/factors/features)
  - Presence of molecular/biological subtypes
  - Various ML or traditional modeling approaches, frameworks, and characteristics

Overview of ML Working Group and Purpose

- This study is not an ML competition, where often the objective is to determine the most accurate method for a particular dataset having outcomes

- This Phase 1 study is not primarily focused on finding new and meaningful biology
  - The patient data is often not “complete”, there are various levels of missing values in certain possible features, there is likely selection bias, there could be incorrect entries or lab test results

- This effort is a Human Learning exercise so that we can make proper recommendations for Clinical Grade best practice AI/ML methods for use in human health and diagnostics

ML Working Group High-level Goals

- Publish results from the study in a notable journal and have these results influence best practice and potentially FDA guidance on ML models used in clinical (and possibly pre-clinical) practice
  - If you participate in a Data Analysis Team submitting results/classifications, you will be listed as a co-author on the main paper
  - Your participation in other facets, including other papers related to the effort, will be judged by the leaders of those papers for co-authorship and leadership
- Make results generally available to the scientific community through the practice of Open Science
  - Implies making the data and the models used to classify them publicly available as well as the model results and the meta-analysis of those models
- Understand the dynamics of ML model evolution for different frameworks and approaches

General Overall Long-Term Objectives for Study and ML Datasets

MAQC Society

- Identify one or more large(r) dataset types (ideally 3 or 3 Phases) with
  - Large number of observations or samples (thousands if not tens of thousands or more)
  - Well-identified classes or outcomes, although having one dataset with confidence scores for outcomes is also welcome
  - Preferred: if we have at least 2-3 subtypes in the datasets
  - Preferred: at least one high dimensional tabular dataset (Phase 1), at least one dataset with extensive molecular characterization of RNA and/or DNA and/or protein (Phase 2), at least one dataset involving clinical imaging (Phase 3)
  - Data characteristics where we can make straightforward subsets of observations, whether by randomization or stratification combined with randomization, for proper held-out testing
- Potential Objective: Limit specialized knowledge related to the particular endpoint in the study so that our results are more likely generalizable to other datasets and endpoints
  - In Phase 1, both for patient identity protection and to promote more generalizability of methods, certain well defined data entries such as particular treatments or drugs (eg) may be tokenized. Other things such as gender, race, cancer stage, diagnostic codes (since there are many and several are related) will probably not be tokenized.
  - In any case, it is not expected for you to be an expert or specialist in a disease area to have meaningful participation in this study, especially for Phase 1.

Overview

- Community participation and setup similar to MAQC-II
  - Steering Committee
  - Data Management Team (Data Prep and Administration Team - one per dataset)
  - ML Analysis Teams (multiple)
  - Meta Analysis Teams (possibly multiple)
  - Potential other teams for later phases: (eg) Federated learning teams
- The steering committee provides oversight for the entire study. Its decisions, achieved via consensus, will be viewed as final.
- Plan for execution to take place in 3-4 months with follow-on analysis of 2-5 months and paper drafted within 1 year of initial data distributions of Phase 1.

MAQC Society

Data Analysis Teams

- Requirements of Data Analysis Teams:
  - All participants must sign NDAs (non-disclosure agreements). All Data Analysis teams must submit a Data Analysis Plan (DAP) using the provided template (at most 2 pages) to the Steering Committee.
    - The Steering Committee will review all Analysis Plans
  - Once you commit to a dataset, you need to provide results for all iterations
  - Highly automated analysis (but not necessarily requiring full automation)
  - Sharing, transparency, and portability of Stat/ML models
    - The Stat/ML models are being measured in addition to the detailed health data
- Suggestions for Data Analysis Teams:
  - Data Analysis Plans include methods for measuring key features of the Stat/ML model(s) implemented (eg, # of nodes, features used, size of implemented code, summary complexity)
- Options for Data Analysis Teams:
  - For Phase 1, you can provide results for one or both datasets – You can choose which dataset
  - You can submit more than one set of classifications (reflecting a distinct ML model) to be evaluated per dataset per iteration. However, you must commit to continuing to evolve that model through all iterations. For example, one model may limit the feature set considered. Another may use a different or competing ML framework.

Detailed Study Goals

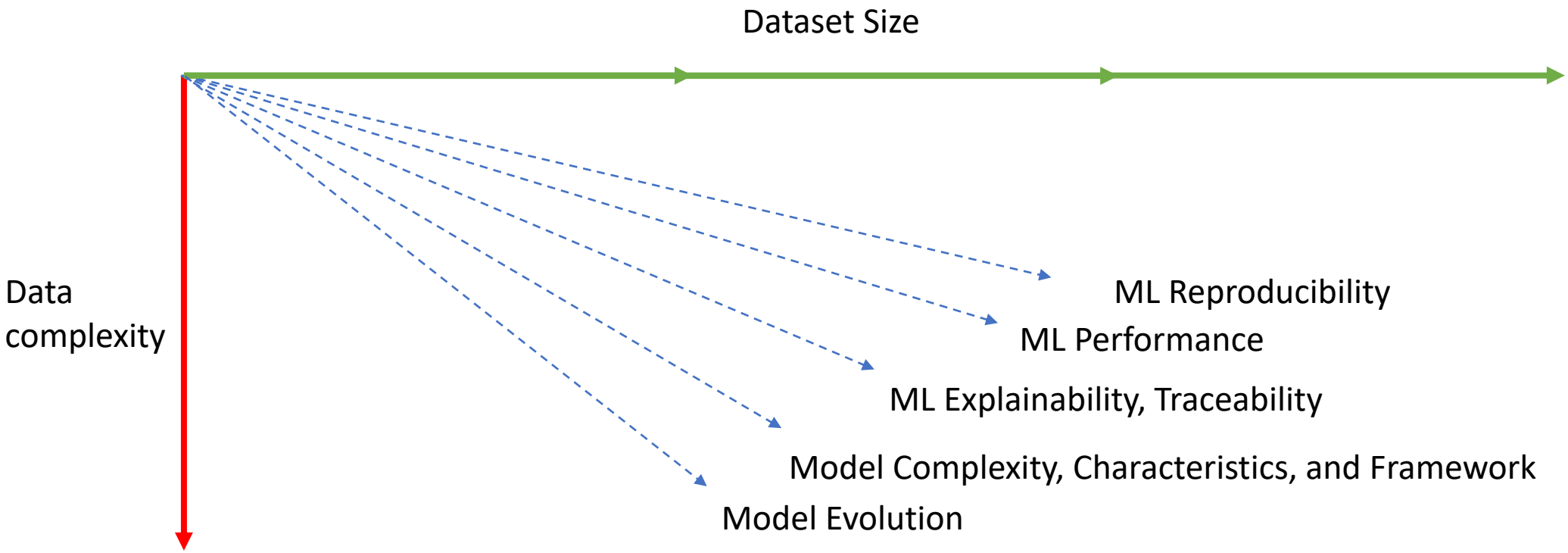
MAQC Society

- Analyze performance, reproducibility, complexity, explainability, traceability of ML and traditional models, at different levels of data availability (dimensions) and **as more data observations are made accessible for training**
  - Track these characteristics with data complexity (ie, correlate)

1. Performance (confusion matrix performance characteristics) – straightforward to compare performance results from models using smaller subsets of data to larger ones
2. Reproducibility – Analyze model performance and characteristics of model structure between cross-validation (CV), fitted values for trained model, and hold-out-sets at different levels of data sizes, between distinct dataset subsets of same size (if pursued), and between dataset subsets of different sizes
3. Complexity - compare against the complexity of the full model and/or full data itself
   - ML-specific – (weighted) number of parameters, nodes, layers, or edges, etc.
   - LM-specific – (weighted) number of covariates or dof, Fisher Information, etc.
   - Data-specific - # of primary factors in PCA or FA, Minimum Description Length (MDL), # of clusters, etc.
   - Class outcomes - Bayes error (if soft labels are available)
4. Explainability – related to ML model evolution
   - In general, one would expect the explanatory factors of models to become more complex as the models become more complex and learn from a larger number of observations. Can this be detected?
   - Constraints – What if we constrained the model complexity? Would it improve explainability? Could we have explainability-related explicit constraints?
5. Traceability – As models evolve, some observations/patients previously classified will change classes. Can we determine and suggest best methods to communicate why?

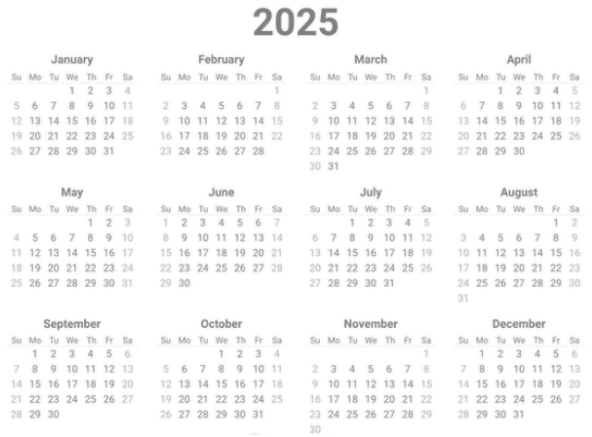
Overview

Examine relationships between



in the context of Human Health Data

General Timeline



MAQC Society

- Kick-off meetings - Feb 13 & 20
- Release of NDAs and DAP template - Feb 21
- Getting NDAs, DAP completed and returned - 1st week Mar
- Release of all feature data (download – roughly 500 MB to 1 GB) - March 10
- Release of first outcome set - March 14
- First training/classification iteration concludes (classifications submitted) - end of March
- Last training/classification iteration concludes - middle of May
- Release of all meta data and instructions for archiving code - end of May
- Beginning of meta analysis – weekly and/or biweekly meetings - June
- Determination of potential papers - first of July
- Paper investigations/ writing - July - Oct
- Preliminary Presentations at 8th Annual Meeting of MAQC Society, Boston - August 11&12
- First drafts available for review - 4Q25
- First submissions - 1Q26

A Few Notes About the Data

- Feature Data will follow an OMOP-like format (Released at one time)
  - Main table with unique entries of patients and straightforward metrics such as
    - Unique patient ID, Gender, Approx. Age (rounded to 5 years), Stage of tumor when diagnosed, Diagnosis Date, Tumor Subtype (if known), region of country (US), Race (sometimes unknown), 3 digit zip
  - Other tables have 0 to multiple entries for each patient related to
    - Tumor molecular test results (varying degrees of missing values) – patients selected if they have ≥ 1 test
    - Blood test results (varying degrees of missing values)
    - Initial or earlier patient therapies (drugs prescribed, etc)
    - Initial or earlier patient treatment procedures
    - Diagnosis codes related to LC/BC and other symptoms described by the patient before and up to diagnosis with LC/BC
    - Detailed tumor stage data at time of diagnosis (Grade, Gleason scores, T value (size))
    - Whether LC/BC is likely primary or metastatic, and if they were diagnosed with any other cancer before and when (relative to diagnosis of LC/BC)
  - Data dictionary will be made available (will not include detailed information like ICD codes)
- Withheld data for classification/prediction will have the Unique Patient ID and the survival value (0/1), released at various points in the study. 5 releases total, 4 training iterations

MAQC Society

Challenges with the Data

- There is likely selection bias associated with the data.
  - We required that all patients have a gender, age, stage, diagnosis date, and an evaluable outcome to the survival threshold chosen.
  - We also required that each patient have at least one molecular test available.
- There will be a large amount of “missing” data. How you choose to deal with missingness will be your decision.
  - For example, there may be a dozen different molecular test results possible for the Lung Cancer patients, but you will see a variable number of results (generally POS or NEG) per patient. There will be many unknowns.
- Patients may have missing information related to therapies and treatments. Remember this is an observational, real-world study.
- The good news is that when we examined survival curves related to primary risk factors, they matched expectations from literature.
- Having said that, this is a reminder that the main purpose of this study is NOT to uncover new biology or novel clinical results. If that happens, that is great, but we should not expect it due to potential selection bias and missing values/incomplete data.
- Any biological or clinical result would require reproducibility with a better study design and follow-up .

Data Challenge Framework

- The WG Data Management team (separate from Data Analysis Teams) will manage the datasets for the MAQC AI/ML Study.
  - In Phase 1, the Data Mgmt team will highly overlap with the Steering Committee
- The WG Data Mgmt team will also prepare various subsets of the outcomes for the purposes of learning, testing, and analysis
  - All feature data for all patients will be released prior to the first set of outcomes (for feature examination, summarization, acclimation)
  - Dataset size for training will evolve (grow) through a series of iterated unveiling of the outcomes
  - Constraints or variations of the basic iterative analysis may be discussed by the group (explorations)

1 Dataset
ML Training Iterations

Sizes of data (number of observations with outcomes)
Where x ≥ 1 and

MAQC Society

Training Size →

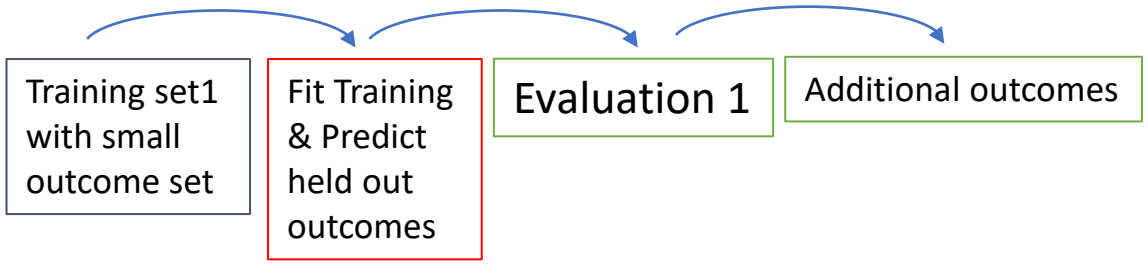
Sample size Small (100x)

Sample size Low (300x)

Sample size Medium (1200x-2400x)

Sample size Large (1000's to 10,000's)

Iteration 1



Steering Committee

Data Analysis Teams

1 Dataset
ML Training Iterations

Sizes of data (number of observations with outcomes)
Where x ≥ 1 and

MAQC Society

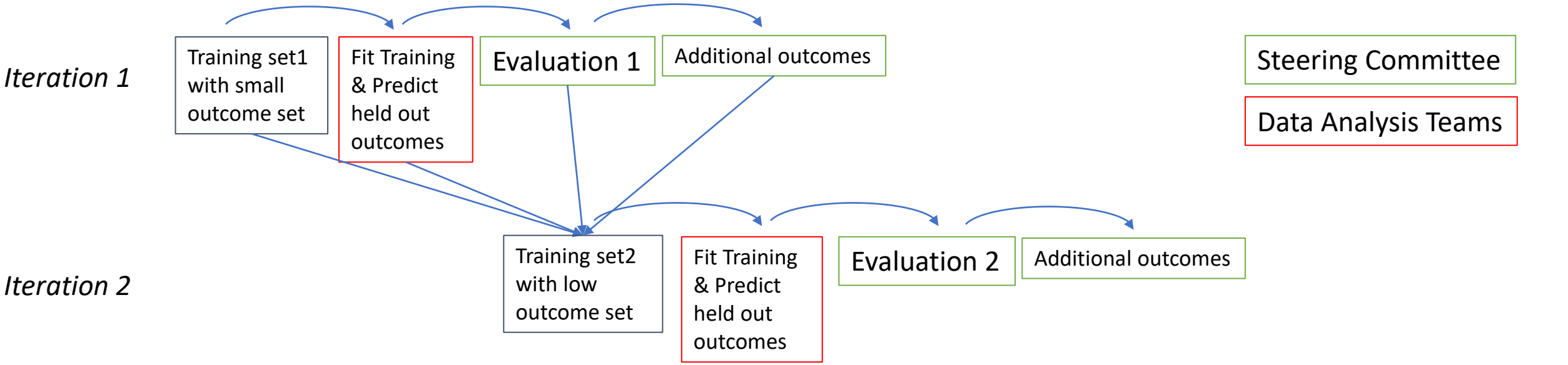
Training Size →

Sample size Small (100x)

Sample size Low (300x)

Sample size Medium (1200x-2400x)

Sample size Large (1000's to 10,000's)



1 Dataset
ML Training Iterations

Sizes of data (number of observations with outcomes)
Where x ≥ 1 and

MAQC Society

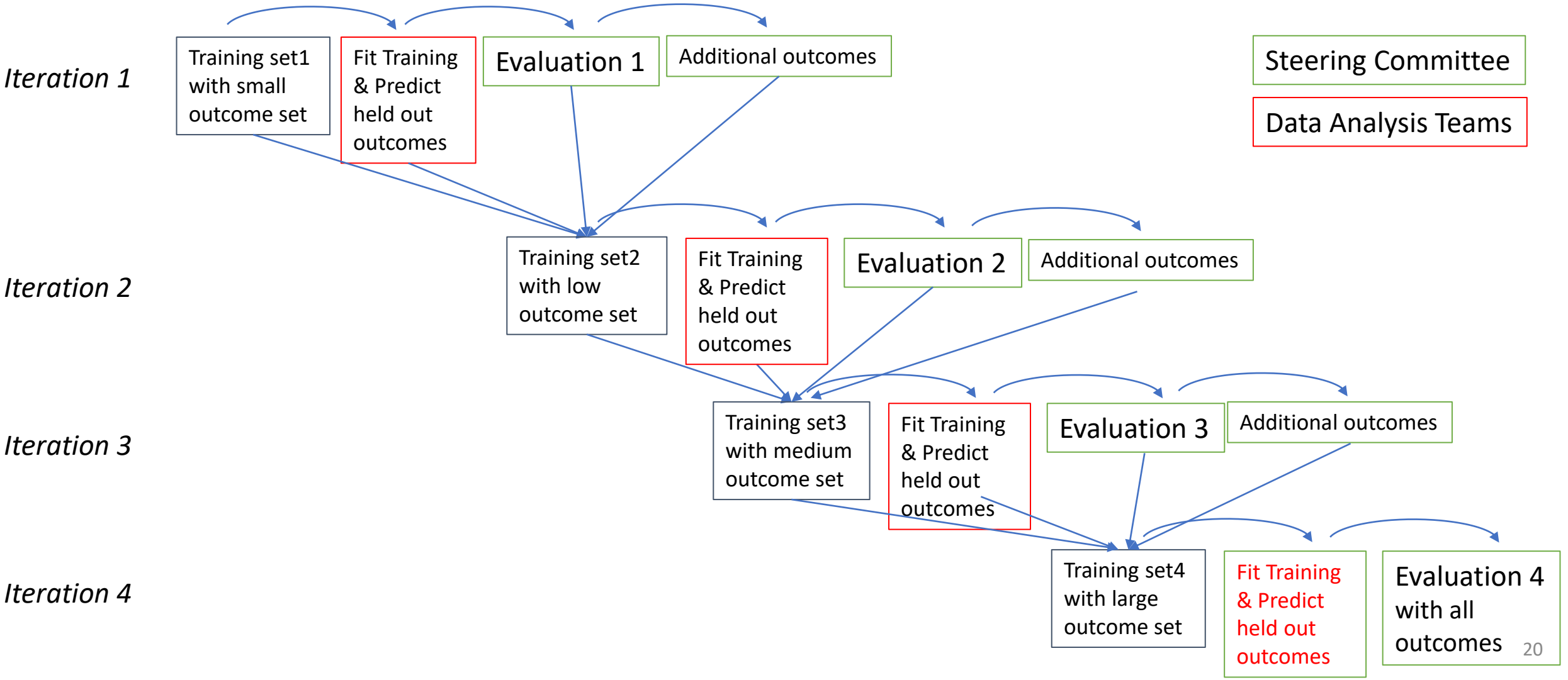
Training Size →

Sample size Small (100x)

Sample size Low (300x)

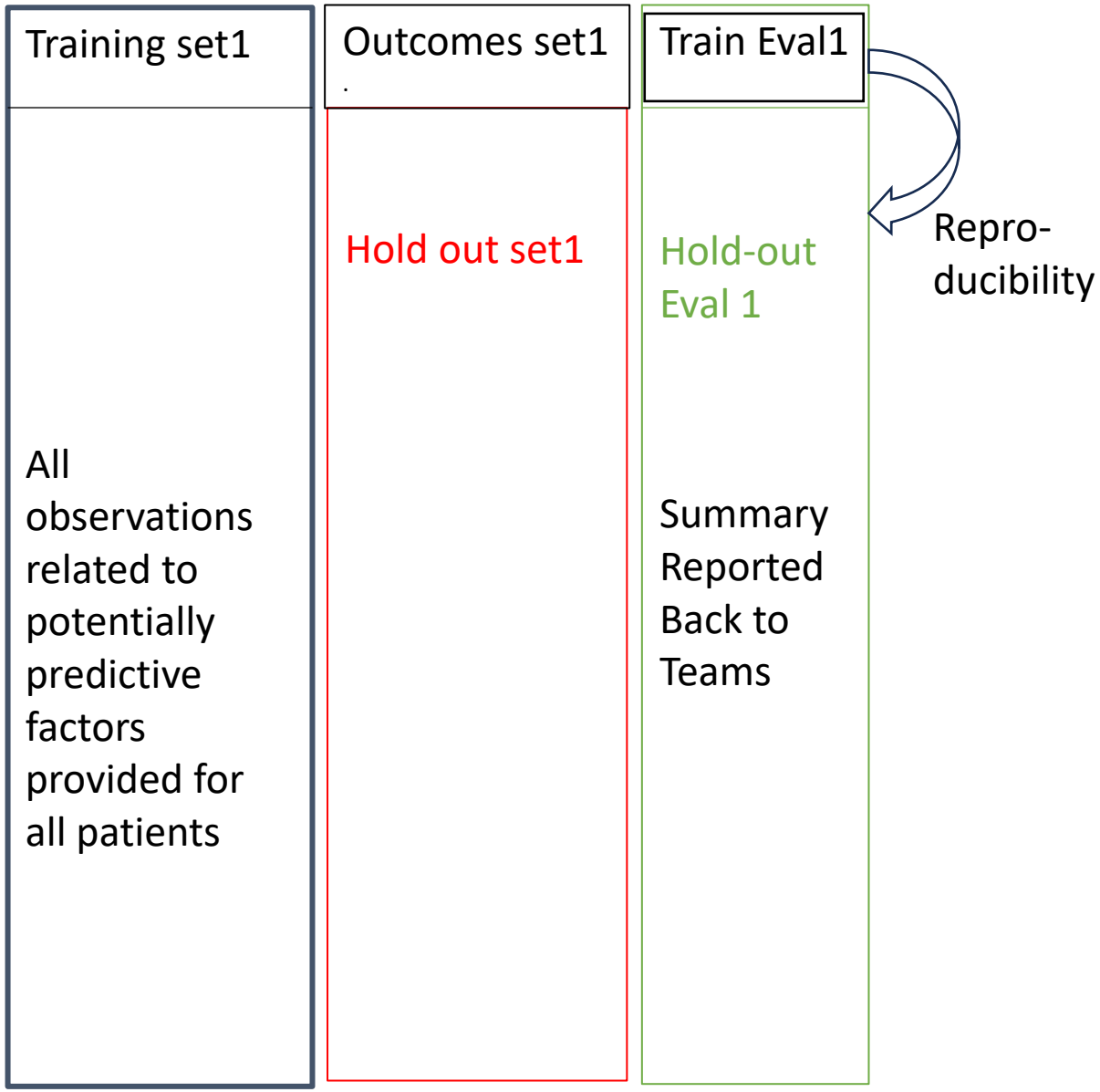
Sample size Medium (1200x-2400x)

Sample size Large (1000's to 10,000's)



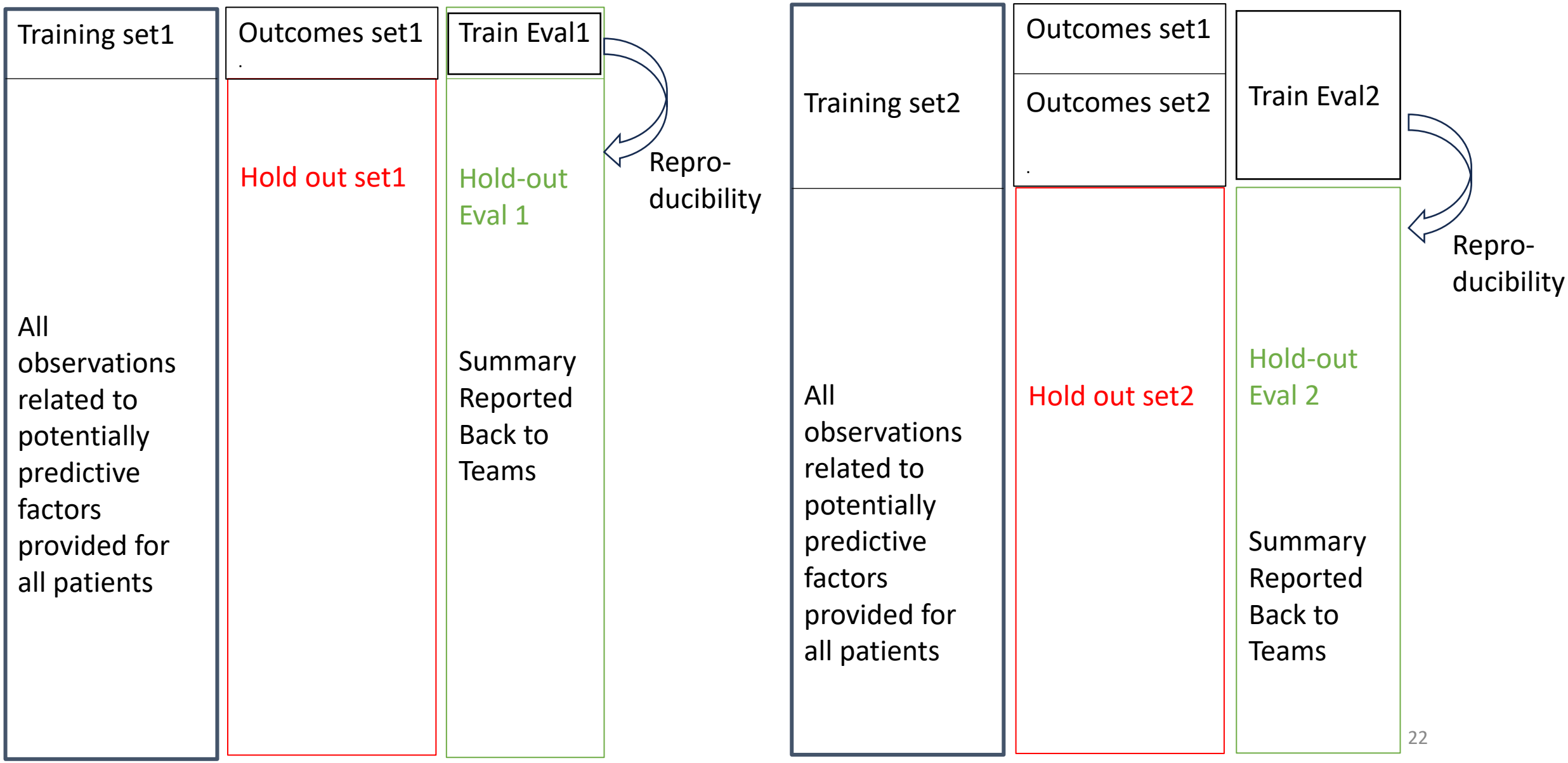
Not Disclosed until later iterations or the end

More Detail on Data Release to Teams



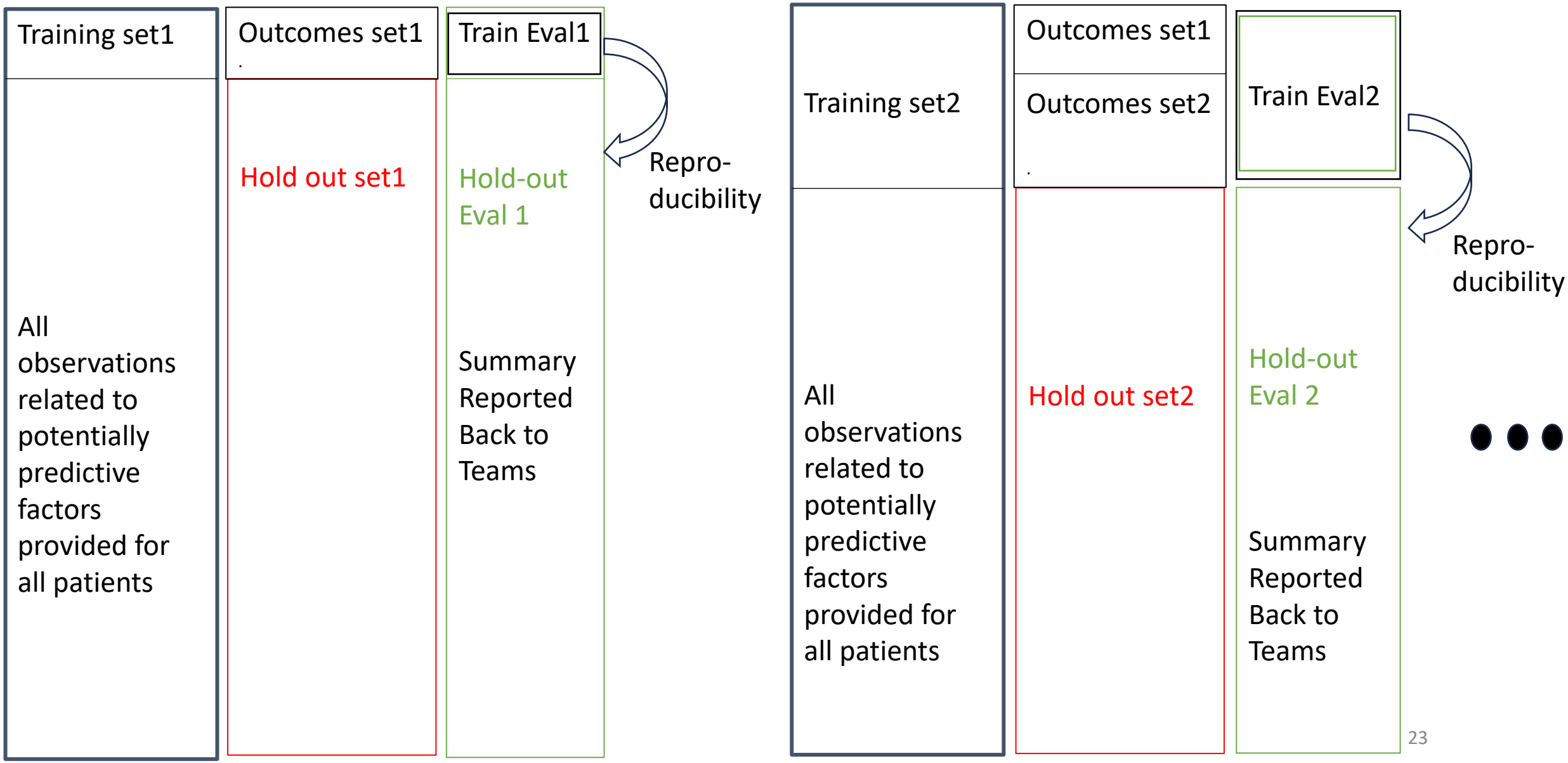
Not Disclosed until later iterations or the end

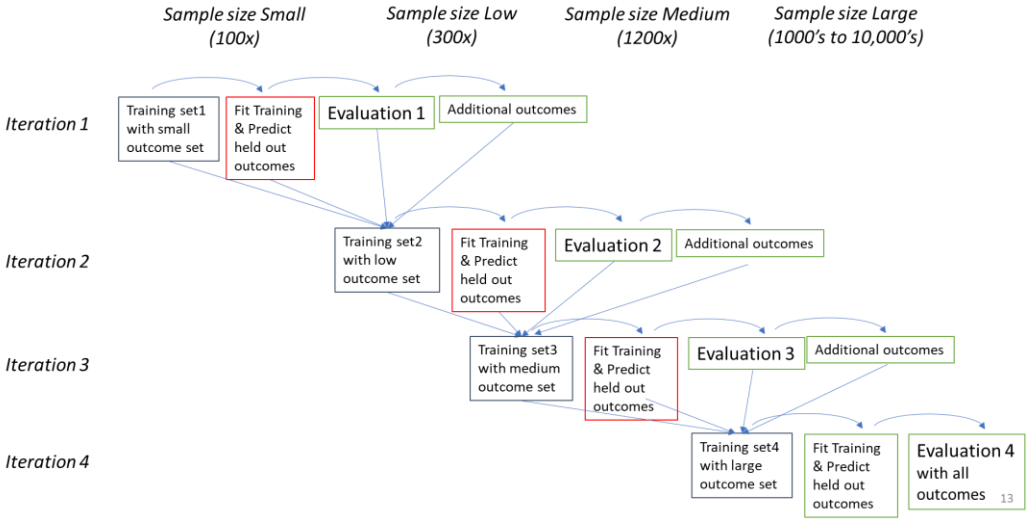
More Detail on Data Release to Teams



Not Disclosed until later iterations or the end

More Detail on Data Release to Teams



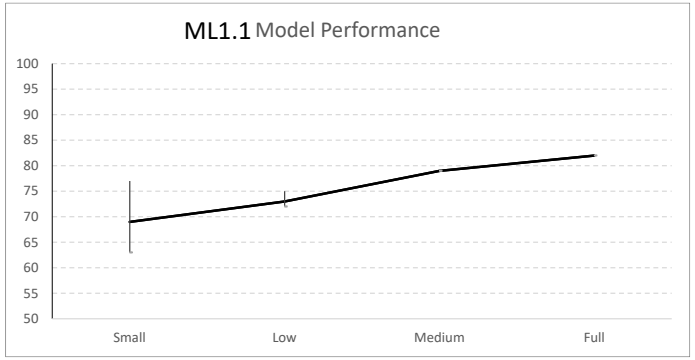


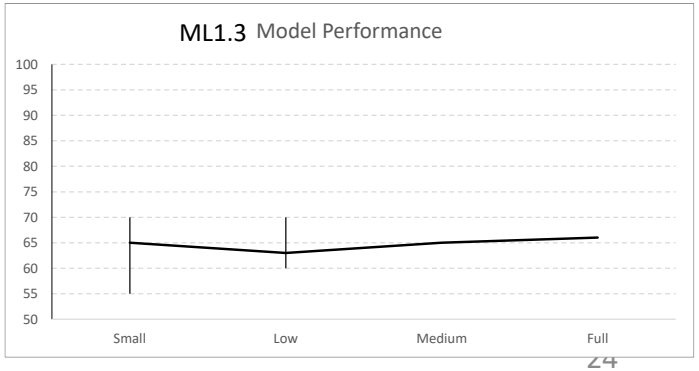
Data Analysis Teams can submit more than one ML/Stat model if they so choose. We ask that the model labeled #1 be the expected "best" model (performant, reproducible, etc.).

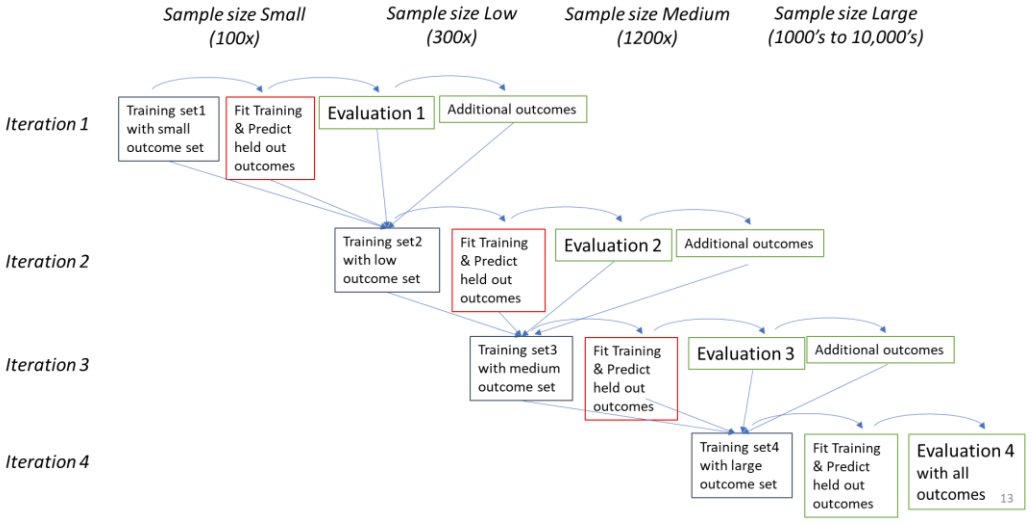
After ML Iterations completed, all ML models and their code/documentations and executables will be made available to all within the study who want to participate in the meta-analysis.

Meta Analysis

Models	Analysis Team	ML or Stat Framework	Complexity	Performance (MCC, F1, Recall, Precision, AUC)	Reclassification Rate	… (eg, Training stats, other summary indicators)
			Iteration 1 – 2 – 3 – 4	Iteration 1 – 2 – 3 - 4	Iteration 2 – 3 - 4	…
ML1.1	1	SVM				…
ML1.2	1	XGBoost				
ML1.3	1	LM (Logistic Reg)				…
ML2.1	2	ANN				…
ML3.1	3	kNN				…
…	…	…	…	…		





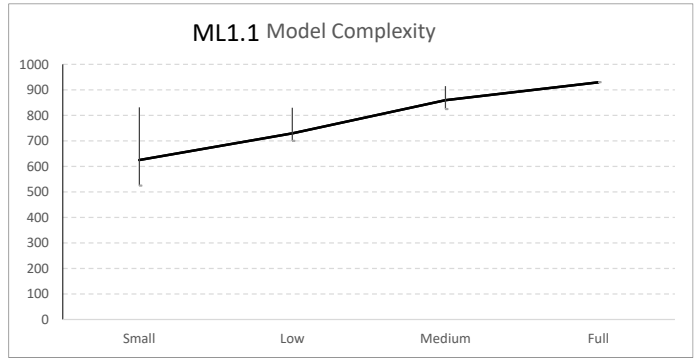


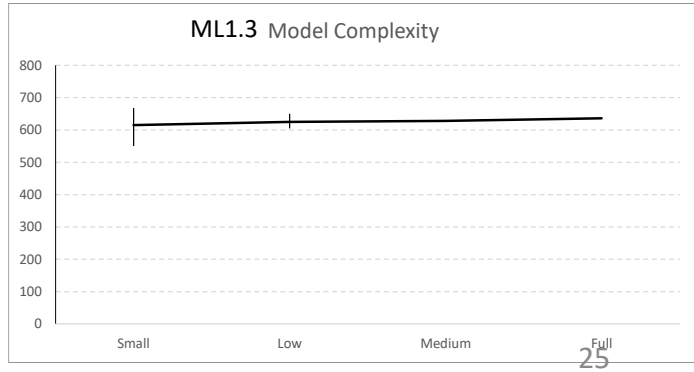
Data Analysis Teams can submit more than one ML/Stat model if they so choose. We ask that the model labeled #1 be the expected "best" model (performant, reproducible, etc.).

After ML Iterations completed, all ML models and their code/documentations and executables will be made available to all within the study who want to participate in the meta-analysis.

Meta Analysis

Models	Analysis Team	ML or Stat Framework	Complexity	Performance (MCC, F1, Recall, Precision, AUC)	Reclassification Rate	… (eg, Training stats, other summary indicators)
			Iteration 1 – 2 – 3 – 4	Iteration 1 – 2 – 3 - 4	Iteration 2 – 3 - 4	…
ML1.1	1	SVM				…
ML1.2	1	XGBoost				
ML1.3	1	LM (Logistic Reg)				…
ML2.1	2	ANN				…
ML3.1	3	kNN				…
…	…	…	…	…		





Submitting Classifications from Iterations

- 1 file per study dataset per model per iteration
- All patient's classifications will be included (some are fitted values from training, some are the held-out test set)
- File type will be a simple tab-delimited text file
  - nomenclature of file name will be communicated
- Simple content (mandatory and optional)

Mandatory Optional

PATIENT_ID	Class	Probability of Class 0	Probability of Class 1
15204128	1	0	1
15204235	1	0	1
15222546	0	0.7	0.3
15319913	0	0.8	0.2
15321157	0	0.55	0.45
15323560	1	0.25	0.75
15324259	0	0.9	0.1
15326575	0	0.99	0.01
15331872	1	0.1	0.9
15336446	1	0.3	0.7
15337169	0	0.7	0.3
15347335	0	0.8	0.2
15364375	0	0.95	0.05
15430473	0	0.95	0.05
15434086	1	0.3	0.7
15458224	1	0.45	0.55

Expected Findings

- Determine reasonable expectations for reproducibility given size, completeness, and complexity of data
  - Level of reproducibility of training vs. test
  - Level of reproducibility of previously classified patients
  - Level of reproducibility of the model framework with independent training
  - Conservation/Mutation of the model framework iteration to iteration
- Better measures of data and model complexity that are informative after-the-fact (after model creation)
- Better understanding of explainability and its relationship to performance and traceability
  - For example, is it reasonable to constrain ourselves to models that offer better explainability and traceability for clinical purposes and traceability?

MAQC Society

Key Concept

- **Explainability** of a ML model is often a general property related to identifying important features, their interactions, and their (relative) impact
  - Features which if not available may greatly negatively impact model performance
- **Traceability** is defined here as
  - The ability to understand why a patient was classified one way or the other. Most important is a situation where with small data training, a patient is classified (eg) 0 and with large data training, the patient is classified not “0”.
  - If models improve performance with larger training data, this WILL occur. Can we explain (to any degree) why an individual patient changed classes?
    - Random chance (patient is at a border) but most borders are stable
    - Subtype emerges (with more data)
    - Model changes more dramatically related to previously sparse data region(s)

Initial Candidate Dataset

- Real-World Evidence (RWE) Electronic Medical Record (EMR) data from IQVIA (CRO and health data science company)
- Breast Cancer - ~236,000 patients (raw) – ~12,000-16,000 that have more complete data
- Lung Cancer - ~80,000 patients (raw) – ~8000-9000 that have more complete data
- Initial Goal was to identify ~10,000 or more patients with more complete data

High Level Data types of RWE Phase 1 dataset

- Demographics – age, gender, pop. group but not behavioral data
- Diagnosis – cancer type, tumor stage, other basic lab results, histology codes, status
- Facility of treatment and lab tests or results
- Treatments – prescriptions/drugs and dose, therapies, radiation, procedures
- Limited -omic information
  - BRAF, EGFR, KRAS, CK7, c-Met, PD-L1 mutation, ALK/ROS1/RET fusion - ER/PR/HER2 status, minority with BRCA sequencing summaries
- Dates – diagnosis date of LC/BC, dates of previous other cancers, other dates to determine/confirm survival information, initial treatments
- Death indicators

High Level Data types of RWE Phase 1 dataset

- Demographics – age, gender, pop. group but not behavioral data
- Diagnosis – cancer type, tumor stage, other basic lab results, histology codes, status
- Facility of treatment and lab tests or results
- Treatments – prescriptions/drugs and dose, therapies, radiation, procedures
- Limited -omic information
  - BRAF, EGFR, KRAS, CK7, c-Met, PD-L1 mutation, ALK/ROS1/RET fusion - ER/PR/HER2 status, minority with BRCA sequencing summaries
- Dates – diagnosis date of LC/BC, dates of previous other cancers, other dates to determine/confirm survival information, initial treatments
- Death indicators

Key Decisions

- Proportion of outcomes released each iteration
  - Would like to follow something similar to Clinical Trial Phase1/2 sizes, Phase 3, 40-50%, 80%
- Overall sample size vs. outcome class proportions (threshold for mortality)
- Use one or both cancer datasets, in parallel or serially
- Timeframe for each training iteration
  - 2 weeks? 3 weeks? 3 weeks for first iteration? Then 2 weeks per after?
- Later decisions
  - Method and location for storing code and ML models (to reproduce classifications for each iteration)
  - Methods for measuring data and model ML complexity
  - Methods for examining explainability and traceability of ML models

First Decision

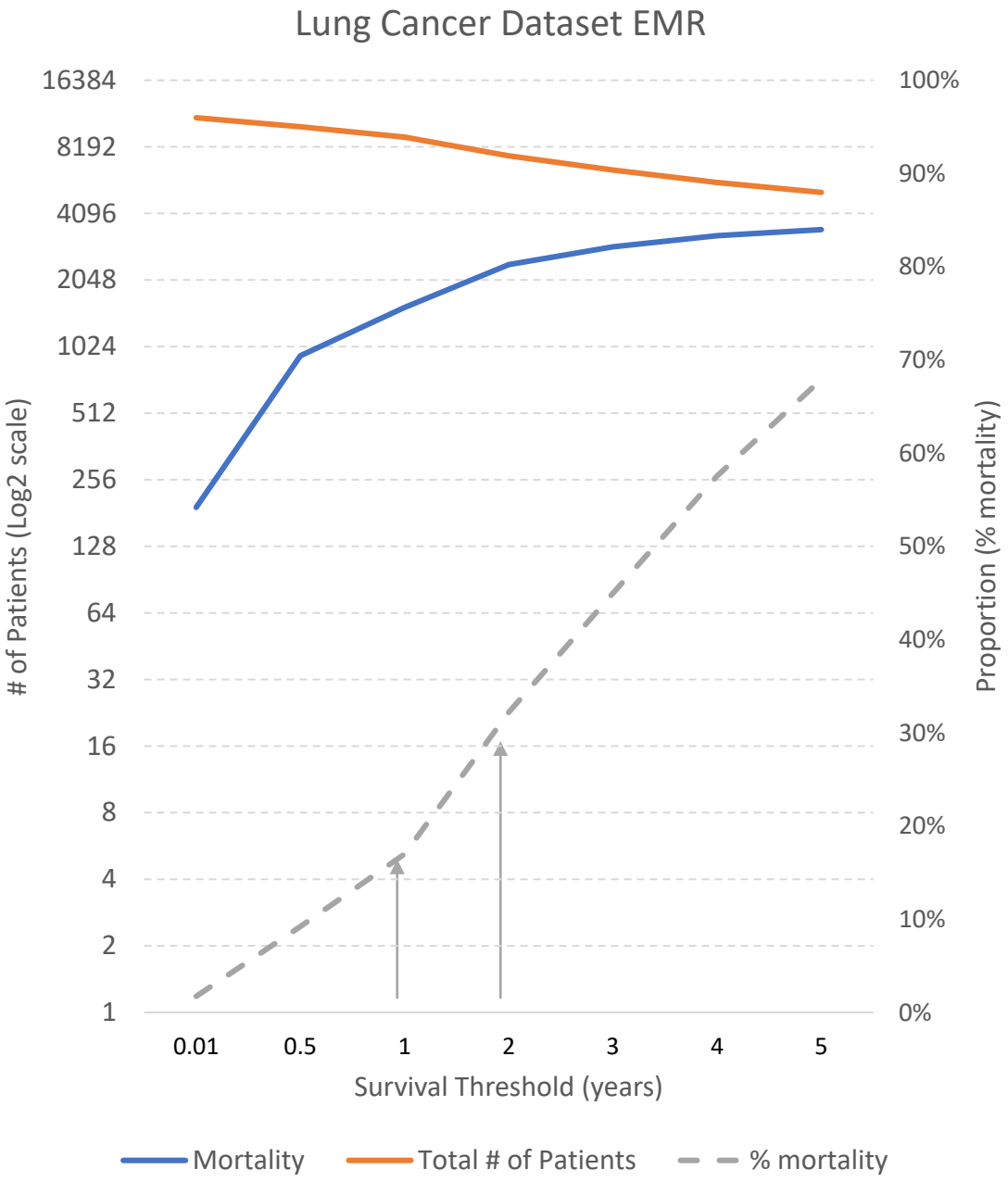
- Proportion of outcomes released each iteration
  1. Recommendation: Would like to follow something similar to Clinical Trial Phase1/2 sizes, Phase 3 sizes, 40-50%, 75-80%
     - 100-150 (2%), 350-500(5%), 40%, 80% (enough so that 500+ patients held-out in last iteration)
  2. 5%, 10%, 40%, 80%
  3. 10%, 25%, 50%, 80%

- The level of missingness of certain factors may make some not as usable until the later iterations
  - Reality of real-world observationally-driven medical data

Second Decision

- Overall Sample Size vs. Outcome Class Proportions (survival threshold)

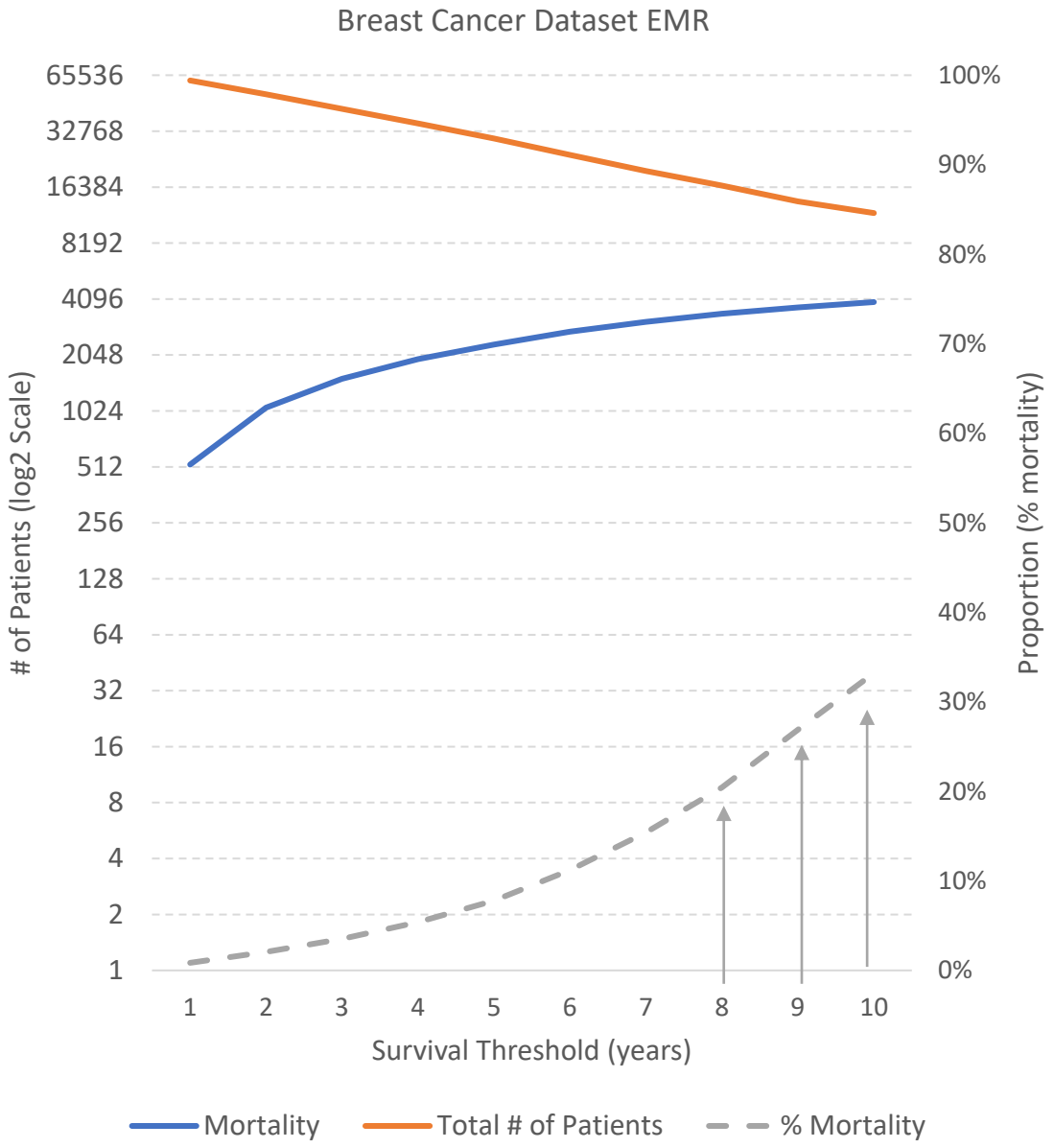
After Year	Deceased by Threshold	Alive at threshold /eventual ly dec.	Alive at threshold	Total Patients	% Deceased
0.01	192	3945	6961	11098	1.73%
0.5	930	3207	5962	10099	9.21%
1	**1540**	**2597**	**4929**	**9066**	**16.99%**
2	**2402**	**1735**	**3322**	**7459**	**32.20%**
3	2894	1243	2299	6436	44.97%
4	3250	887	1511	5648	57.54%
5	3461	676	960	5097	67.90%



Second Decision

- Overall Sample Size vs. Outcome Class Proportions

After Year	Deceased by Threshold	Alive at threshold /eventually dec.	Alive at threshold	Total Patients	% Deceased
1	527	4348	56533	61408	0.86%
2	1067	3808	46861	51736	2.06%
3	1526	3349	38363	43238	3.53%
4	1945	2930	31170	36045	5.40%
5	2332	2543	25069	29944	7.79%
6	2733	2142	19529	24404	11.20%
7	3085	1790	15124	19999	15.43%
8	**3412**	**1463**	**11818**	**16693**	**20.44%**
9	**3692**	**1183**	**8858**	**13733**	**26.88%**
10	**3945**	**930**	**7002**	**11877**	**33.22%**



Third Decision

- Use only one dataset, both datasets?
  - Main advantage to LC is a bit more variety and extent of molecular markers and more variety in gender (obviously)
  - Main advantage to BC is a larger sample size (almost 2x) and more likely for tumors to be primary. Also, there is a very high percentage of patients who have results for the 3 primary assays associated with BC diagnosis (ER/PR/HER2)
- If both, then analyze serially or in parallel?

Fourth Decision

- Timeframe for iterations
  - 3 week intervals between data release/training & submissions
  - 3 week interval for first data release/training to allow for time to acclimate to data, 2 week intervals after that (3-2-2-2)
  - ~2 week intervals between data release/training & submissions

Summary

- There are many challenges to making AI/ML models and results to be Clinical Grade so that they improve human health
- Ideally, these models and results should be robust, stable, reproducible, explainable, and traceable
- What can we as a society and a collaborative community do to assist and promote this general challenge and opportunity?
- The MAQC Society is an open society. If you have colleagues who you think would like to participate, there is still time to join the effort.

Poll Results

			2/13/2025	2/20/2025								
Polls			**Option 1 Meeting**	**Option 2 Meeting**								
	First Decision		**Proportion of Outcomes released at each iteration**									
			40%	50%	1. 2%- 5% - 40% - 80% (Clinical Trials + RWE)							
			5%	17%	2. 5% - 10% - 40 % - 80%							
			55%	33%	3. 10% - 25% - 50% - 80%							
	Second Decision (a)		**Setting Survival Threshold - Lung Cancer (sample size/class balance trade-off)**									
			35%	13%	1. Threshold at 1 year							
			55%	75%	2. Threshold at 2 years							
			10%	13%	3. Threshold at 3 years or higher							
	Second Decision (b)		**Setting Survival Threshold - Breast Cancer (sample size/class balance trade-off)**									
			40%	70%	1. Threshold at 8 years							
			20%	0%	2. Threshold at 9 years							
			40%	30%	3. Threshold at 10 years							
	Third Decision		**Dataset Scope of Phase 1 (Multiple Choice)**									
			20%	15%	1. Lung Cancer only							
			5%	0%	2. Breast Cancer only							
			75%	85%	3. Both Breast and Lung Cancer							
					if both, then							
			75%	50%	1. Serially							
			25%	50%	2. Parallel							
	Fourth Decision		**Time intervals for Submission (Training Cycle + submission)**									
			40%	50%	1. 3 week intervals between release of data and submissions							
			55%	50%	2. 3 week interval for the first submission and then 2 weeks per iteration after							
					3. 2 week intervals between release of data and submissions							
			5%		4. Something else							
	Attendance		20	14								