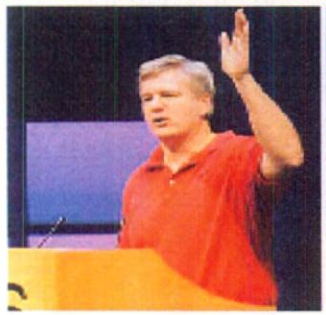


Russ Wolfinger, SAS Director of Scientific Discovery and Genomics, chairs the MAQC meeting.

John Sall, SAS Co-founder, welcomes MAQC meeting participants.

FDA's Bob Wagner, Reena Philip, Uwe Scherf, and Federico Goodsaid discuss "FDA Regulatory Perspectives on Genomics".

It is how the MAQC meeting looks like; everyone is listening carefully. Who is the amazing presenter? ☺

MAQC-II Review of Data Analysis Protocols (DAPs)

Date: Wednesday, January 9, 2008
Time: 6 am PST / 8 am CST / 9 am EST / 2 pm GMT
Duration: Up to 5 hours, if needed

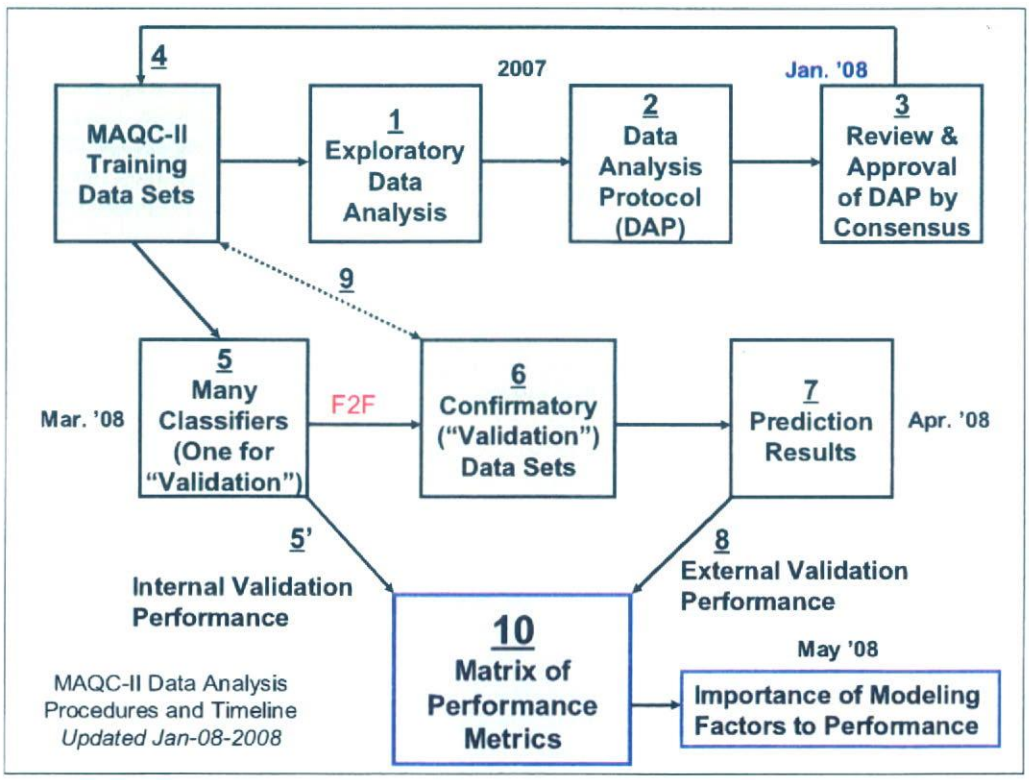
Audio:
Dial-in Number: 1-866-296-6844 (Toll-free; US only)
+1-816-249-4809 (International)
Conference Code: 8549731585

WebEx:
Meeting Number: 333 006 275
Meeting Password: MAQC-II

To join the online meeting:
1. Go to https://rosettabio.webex.com/rosettabio/j.php?ED=95847732&UID=56590852
2. Enter your name and email address.
3. Enter the meeting password: MAQC-II
4. Click "Join".

We are grateful to Dr. Mette Peters (Mette_Peters@rosettabio.com) of Rosetta Biosoftware for providing the TC and WebEx bandwidth to the MAQC-II.

Leming.Shi@fda.hhs.gov (Tel: +1-870-543-7387)



Agenda

1. Overview of the Agenda (Leming Shi, 5 mins)
2. DAP submission and review processes (RBWG – Greg/Tim/Lakshmi, 10 mins)
3. Survey of the 33 DAPs (Attachment #1), revisit of MAQC-II objectives (Attachment #2), and thoughts and suggestions (Leming Shi, 30 mins)
4. General discussion on the DAPs and the review comments (All participants, 30 mins)
5. Discussion on the DAPs from six European teams (DAT# listed in Attachment #3):
   Spheromics – Trygg Bylesjo Scherer (#25)
   Almac – Jurergen Frese (#35)
   CIPF – Joaquin Dopazo (#6)
   DKFZ – Benedikt Brors (#8)
   FBK – Cesare Furlanello (#14)
   SIB – Vlad Popovici (#24)
6. Discussion on the DAPs from four Chinese teams:
   CBC – Liang Zhang (#1)
   CAS – Tieliu Shi (#2)
   Tsinghua – Xuegong Zhang (#27)
   ZJU – Xiaohui Fan (#36)
7. Discussion on the DAPs from 18 US teams:
   According to the order shown in the December 06, 2007 list of Data Analysis Teams (DATs) – Attachment #3
8. Next steps
   Update your DAP and submit it to Leming.Shi@fda.hhs.gov; not to be further reviewed.
   Apply your DAP to the data sets
   Report models to MAQC-II (template to be developed and provided soon)
   Observe the timeline listed on the first page of this document
   Plan to attend the next MAQC face-to-face meeting
9. Other items

Attachments:

1. "Survey of MAQC-II Data Analysis Protocols (DAPs)", January 8, 2008 (Leming Shi)
2. "Considerations on the MAQC-II Data Analysis Process", July 10, 2007 (Leming Shi)
3. "Data Analysis Teams (DATs) in the MAQC-II", December 6, 2007 (Leming Shi)
4. "The Matthews Correlation Coefficient (MCC) as a Possible Performance Metric for Assessing the Quality of Classifiers Being Developed by the MAQC Project", September 13, 2007 (Huixiao Hong and Leming Shi)

"***If a Statistical Analysis Plan (SAP) is hardwired to a specific data set, we will more likely end up fitting to noise.***" Dr. Kenneth Hess (M.D. Anderson Cancer Center), MAQC-7 @ SAS

Thoughts and Suggestions from Leming Shi

The purpose of the DAP is not to produce the most "accurate" classifier for a particular data set, but rather to evaluate the general data analysis workflow and to determine whether the workflow works reasonably well on different data sets (endpoints); that is why we are working on multiple data sets (endpoints). It is expected that some workflows will be more robust (work on most data sets), whereas other workflows may be more prone to overfitting and may not extrapolate well to external validation data sets.

An important goal of MAQC-II is to determine the relative importance of modeling factors to the model's performance so that a general data analysis procedure that is likely to work outside of the MAQC-II data sets may be identified.

We have spent a lot of time over the past several months on almost every issue in the model development/validation process. I feel that it is time to make some "tough" decisions so that we can move on. The following are some examples:

1. Performance metric: Matthews Correlation Coefficient (MCC) – Attachment #4. Other metrics can also be provided, but the MCC will be used as the primary metric across all the data sets (endpoints).
2. Normalization/summary methods: While single-array normalization/summary methods are preferred for practical reasons, multiple-array normalization/summary methods such as RMA should be allowed as long as the training set and the confirmatory ("validation") set are normalized/summarized separately, or the training data remain unchanged when the confirmatory data are to be normalized/summarized together with the training data.
3. One versus multiple models per data set (endpoint) per Data Analysis Team (DAT): To address multiplicity issue, it is important for each (DAT) to provide one and only one "best" model for each data set (endpoint) for "validation" purpose. Meanwhile, it is of critical importance for each DAT to also report all the models that are being explored as a result of the many different combinations of the modeling factors described in the DAP. These "research" models will be essential for us to delineate the relative importance of the many modeling factors on a model's performance. However, such "research" models should not be considered as "validated" by the confirmatory data set even if they may perform much better than the chosen "best" model. If the "best" model performs the best in predicting the confirmatory data, we should be happy. However, if it performs much worse than the "research" models, we need to find out why and learn something. Each DAP should clearly state the total number of models to be explored per data set (endpoint).
4. One "best" model per data set (endpoint) for the MAQC-II: Before confirmatory ("validation") data sets can be distributed, the MAQC-II will need to pick its "best" model for each data set (endpoint) to avoid multiplicity issue since the MAQC-II is one single team. A face-to-face meeting should be very helpful to make such decisions.
5. Internal validation performance needs to be reported: It will allow us to examine the degree by which the model's performance may be overestimated in external validation.
6. Model reporting template: Volunteers are needed to develop a template for reporting models.
7. Executables for independent prediction: Each DAT should plan to make its model prediction in an easy to use form by a third party for independent prediction.
8. The 8th MAQC face-to-face meeting: I propose that we have the next (8th) face-to-face MAQC meeting in late March or early April, 2008 when models will have been generated and reported to the MAQC-II. Things to do: Reporting models; Selecting MAQC-II's "best" models; Distributing validation sets; Developing MAQC-II's "best" practice for model development and validation; Identifying additional data sets to test/validate the "best" practice.

Survey of MAQC-II Data Analysis Protocols (DAPs)

8-Jan-08

Leming.Shi@fda.hhs.gov (Tel: +1-870-543-7387)

33 DAPs from 28 Data Analysis Teams (DATs) from 8 countries (US, China, Finland, Germany, Italy, Spain, Sweden, and UK)

DAT#	No.	DAP-Organization	Classification Algorithms	Summary & Normalization	Internal Validation	Performance Metrics	Batch Removal	Feature Filtering	Feature Selection
1	1	CBC DAP - Liang Zhang.doc	KNN, LDA, SVM	dCHIP	10-CV	MCC	Yes	Intensity (75%), FC>2 and P<0.05	
2	2	CAS DAP - Tieliu Shi I.doc	RF, SMO, BN, SVM	gcRMA	10-CV	AUC	L/S (location/scale)	DEDS; absent in >80% samples	Sparse LR, PPI
2	3	CAS DAP - Tieliu Shi II.doc	SVM	MAS5, gcRMA		Acc	Yes	ANOVA	PK-RFE
2	4	CAS DAP - Tieliu Shi III.doc	Linear Logistic Classifier	MAS5, gcRMA	5-CV	Acc, Sen, Spe	Yes	FC>2 (1.5), P<0.05	Correlation
4	5	CDRH DAP - Gene Pennello.doc	KNN, DQDA, DLDA, PCA+Logistic/Cox regression, LASSO	MAS5, Custom	10-CV	AUC, gAUC, # of features, Biosignificance	Ref norm.	Intensity, FC, P	
5	6	Wagner et al. Plan for MAQC2.doc							
6	7	CIPF DAP - Joaquin Dopazo.doc	KNN, DLDA, PAM, SOM, SVM	RMA, Quantile, Lowess	LOO	Acc		F, Wilcoxon	
7	8	ICB DAP baseline - Fabien Campagne.doc (4)	SVM	MAS5				T, FC, GA	k=50 (with largest SVM weights), Pathway/Literature
8	9	DKFZ DAP - Benedikt Brors.doc	PAM, SVM-RFE, SVMsemi-sup.	VSN2, gcRMA	10-CV	Acc	No	No	Embedded
11	10	EPA DAP - Richard Judson.doc	SVM (different kernels), ANN, RF, KNN, Logitboost, PAM, LDA, RPART, NB, J4.8	RMA, gcRMA	10-CV		dCHIP, ComBat	Limma, SAM, PCA	
13	11	GeneGo DAP - Weiwei Shi.doc	DLDA, RF	RMA, MAS5	10-CV	Geometric mean of Acc-Sen-Spe	Yes	No	FC, disease network/pathway
14	12	FBK-MPBA DAP - Cesare Furlanello.doc	KNN, SVM, Tree, Ensemble	MAS5	K-CV	Acc	Standardization	No	Embedded
16	13	Ligand Pharma DAP - Wen Luo.doc	PAM	MAS5	10-CV	Acc	Differential means	Intensity, correlation	
17	14	NCTR DAP - Weida Tong.doc	KNN, BN, SVM, DF	MAS5	5-CV	MCC, Acc			FC (P)
18	15	NIEHS Chou Bushel DAP - Pierre Bushel I.doc	SVM	SVN	10-CV	Acc	Mean-centering and scaling	FC, T	

18	16	NIEHS DAP - Jennifer Fostel.doc	HCA, K-means	RMA					
18	17	NIEHS Li Bushel DAP - Pierre Bushel II.doc	KNN, SVM	RMA	10-CV, LOO	Acc	?	ANOVA	
19	18	NWU DAP - Simon Lin.doc	KNN, SVM, RF	RMA, VSN	5-CV	Acc		Absent, Intensity, P<0.05	P<0.05, Text mining (GeneRIF)
20	19	Princeton DAP - Jianqing Fan.doc	NSC, FAIR, ext.SIS	MAS5, Fan-Niu, SLIM	10-CV				T
21	20	Roche DAP - Mark Fielden.doc	SVM, PAM, LDA, QDA, LR, NN, Tree, KNN, RF	MAS5	10-CV		No	Variance	Many
22	21	SAI_DAP - John Zhang.pdf	SVM, KNN, LDA, QDA, LR, CART, RF	MAS5, RMA	k-CV, LOO, Random split	Acc, Sen, Spe, AUC, RMSE, MCC, Reprodicibility of feature list	Reference batch	Intensity	FC, T, SAM, Welch T, Wilcoxon, ...Consensus
23	22	SAS DAP - Russ Wolfinger.doc	DA, DS (DD), GLM, KNN, LR, Tree, PLS, SVM (RBM)	Mean of PM	5-CV	AUC, RMSE		T, GA	
24	23	SIB DAP - Vlad Popovici.doc	SVM, LDA	MAS5, RMA, VSN	5-CV	AUC, Acc, Sen, Spe	No	Controls, SD<0.75	RFE, Ratio of between-group and within-group sum of squares
25	24	Trygg Bylesjo Scherer DAP - Andreas Scherer.doc	OPLS, K-OPLS	MAS5, RMA	MCCV	Acc, MCC	OPLS, Mean-centering	Intensity	
26	25	SA DAP - Guozhen Liu.doc	KNN		LOO, 5-CV	Sen+Spe	Mean-centering and scaling	Intensity, FC	P
27	26	Tsinghua University DAP - Xuegong Zhang.doc	SVM, RF, BN, KNN	dCHIP, MAS5	LOO	MCC, AUC		Low quality, small overall variability	FC+statistic test, R-SVM, SVM-RFE
29	27	Cedars-Sinai UCLA DAP - Xutao Deng.doc	KNN, BN	MAS5			BEF (Batch Effect Filter)	Intensity, FC (P<0.05)	GA
30	28	UIUC DAP - SHeng Zhong.doc	Tree, BN, SVM, R-SVM, Zhong	MAS5, PLIER	K-CV	AUC, Sen+Spe			Embedded
31	29	UML DAP - Dalila Megherbi.doc	Tree, K-means, SVM, Hybrid	MAS5		Acc, AUC		FC (P), Golub	
32	30	USM DAP - Venkata Thodima.doc	LibSVM, SVM-OBO, SVM-CS, MLP	MAS5	K-CV, LOO	Acc, AUC		FC (P<0.05)	SNR, BW, KW, Gain Ratio
35	31	Almac DAP - Juergen Frese.doc	PLS-DA	dCHIP	LOO, 10-CV	Acc, RMSE	Mean-centering for each variable	~50% reduction (intensity variation <200)	Recursive elimination in CV (smallest absolute weights)

36	32	ZJU DAP - Xiaohui Fan.doc	Nearest-Centroid Classifier (MammaPrint)	MAS5	LOO	Acc, MCC, Sen, Spe	Yes	FC (P<0.01)	
37	33	JHSPH DAP - Rafael Irizarry.doc	Barcode	RMA	K-CV	Acc			

106	50 classification mehtods used 106 times
16	**SVM**
14	KNN
7	RF
5	BN
5	LDA
5	PAM
5	Tree
3	DLDA
3	LR
2	K-means
2	QDA
1	ANN
1	Barcode
1	CART
1	DA
1	DF
1	DQDA
1	DS (DD)
1	Ensemble
1	ext.SIS
1	FAIR
1	GLM
1	HCA
1	Hybrid
1	J4.8
1	K-OPLS
1	LASSO
1	**LibSVM**
1	Linear Logistic Classifier
1	Logitboost
1	MLP
1	NB
1	Nearest-Centroid Classifier (MammaPrint)
1	NN

Summary & normalization	48
MAS5	19
RMA	10
gcRMA	5
dCHIP	3
VSN	2
Custom	1
Fan-Niu	1
Lowess	1
Mean of PM	1
PLIER	1
Quantile	1
SLIM	1
SVN	1
VSN2	1

33	Int. validation
12	10-CV
8	LOO
6	5-CV
5	K-CV
1	MCCV
1	Random split

Performance metrics	51	
Acc	18	
AUC	9	
MCC	6	
Sen	4	
Spe	4	
RMSE	3	
Sen+Spe	2	
# of features	1	
Biosignificance	1	
gAUC	1	
of ACC-Sen-Spe	1	Geometric mean of ACC-Sen-Spe
ility of feature list	1	Reprodicibility of feature list

1 NSC
1 OPLS
1 PCA+Logistic/Cox regression
1 PLS
1 PLS-DA
1 RPART
1 **R-SVM**
1 SMO
1 SOM
1 **SVM (different kernels)**
1 **SVM (RBM)**
1 **SVM-CS**
1 **SVM-OBO**
1 **SVM-RFE**
1 **SVMsemi-sup.**
1 Zhong

Abbreviations:
BN: Bayes Network
DEDS: Differential Expression via Distance Synthesis (FC, T, SAM)
FAIR: Feature Annealed Independence Rules
FC: Fold Change
K-OPLS: Kernel Orthogonal Projections to Latent Structures
LR: Logistic Regression
MLP: Multi Layer Perceptron
NSC: Nearest Shrunken Centriod
OPLS: Orthogonal Projections to Latent Structures
PK-RFE: Polynomial Kernel SVM Recursive Feature Elimination
PPI: Protein-Protein Interaction
SIS: Sure Independence Screening
SMO: Sequential Minimal Opimization (for SVM)

Considerations on the MAQC-II Data Analysis Process

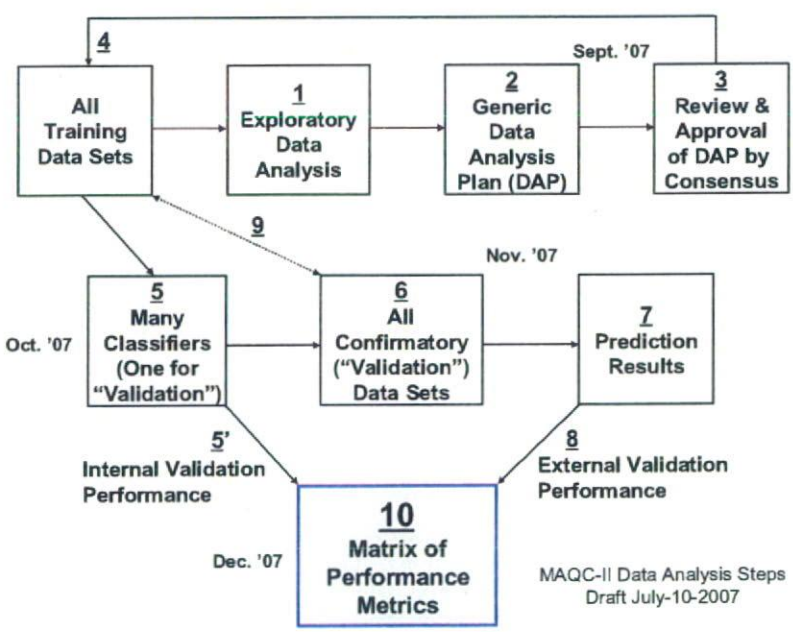
Leming.Shi@fda.hhs.gov July 10, 2007

- Development of Generic Data Analysis Plans
- Review of Data Analysis Plans
- Release of Confirmatory ("Validation") Data Sets
- "One Team, One Plan, for All Data Sets"

"***If a Statistical Analysis Plan (SAP) is hardwired to a specific data set, we will more likely end up fitting to noise.***" Dr. Kenneth Hess (M.D. Anderson Cancer Center)

Key Steps in MAQC-II Data Analysis

1. **Explore all training data sets:** Learn "best practices" for model development.
2. **Develop a generic data analysis plan (DAP):** The DAP should not be "hardwired" to a specific data set; otherwise, we will more likely end up over-fitting.
3. **Review and approve the generic DAP:** Consensus is needed through open discussions.
4. **Apply the generic DAP to all training data sets:** Hard to conclude with only one data set.
5. **Report classifiers and performance metrics (internal validation) to WG co-chairs:** Many classifiers; one to be highlighted for "validation".
6. **Distribute confirmatory (validation) data sets:** Microarray data only.
7. **Report prediction results from all classifiers to WG co-chairs**: 0s and 1s.
8. **Calculate classifier performance metrics (external validation) by WG co-chairs.**
9. **Switch training and confirmatory data sets**: Repeat Steps 4-8.
10. **Meta-analysis of the "matrix of performance metrics":** Determine the relative impact of different factors on model performance.



Our Challenges - Too many factors/options for the development of predictive models: In his presentation on "Predictive Modeling Analysis Factors" during the 7th MAQC meeting at SAS, Dr. Russ Wolfinger correctly pointed out that "we're wandering through a fairly complex space and a large number of dimensions." Russ showed a list of factors and options that might be considered during predictive modeling (see pages 4 and 5). The number of combinations is already extraordinary even without considering your own "favorite" options for these factors. The challenge to the MAQC-II data analysis is how to cover the millions of combinations. While SAS Institute is using a Design of Experiments (DOE) approach to derive a subset of combinations that cover the space, most data analysis teams will most likely be able to explore only a few of combinations of the factor space.

Multiple Types of Data Sets: One fundamental objective of the MAQC-II project is to investigate the impact of different factors on the performance of predictive models (classifiers). However, we cannot draw any concrete conclusions by simply examining only one data set, where "sample size" would be one (in terms of the number of data sets). That is why we have been working very hard to solicit and decide on multiple data sets under the MAQC-II (Table 1).

Table 1. Summary of Data Sets Being Analyzed by MAQC-II

No.	Disease/Study Type	Endpoint	Training Set	Validation Set
1	Hamner TGx	Lung Tumor	Hamner	Hamner
2	Iconix-EPA TGx	Liver Tumor	Iconix	Iconix, EPA
3	NIEHS TGx	Liver Toxicity	NIEHS*	?
4	Breast Cancer	Treatment Outcome; Prognosis	MDACC	MDACC Institut Jules Bordet
5	Multiple Myeloma	Prognosis; Treatment Outcome; Subtypes	UAMS	UAMS, Millennium Univ. Heidelberg Univ. Milan Univ. Hospital Montpellier
6	Neuroblastoma	Prognosis; Subtypes	Univ. Cologne	Univ. Cologne

* The only training data set not yet distributed to data analysis teams.

One Generic Data Analysis Plan for All Data Sets: Applying one data analysis procedure on one data set and another data analysis procedure on another data set will not help us draw any solid conclusions. That is why I have been advocating the idea for each data analysis team to first explore all the MAQC-II training data sets and then develop a generic Data Analysis Procedure (DAP) that will be reviewed, modified, approved, and then applied to all the MAQC-II training data sets. I strongly believe that it is essential to implement this procedure in MAQC-II data analyses; otherwise, we will not be able to populate the "matrix of performance metrics" illustrated as Table 5 in the MAQC-II Research Plan (copied here on page 5 for your convenience). The "one team, one plan, for all data sets" proposal will also greatly help the process of the review of analysis plans.

I would like to bring your attention to a quote from Dr. Kenneth Hess (M.D. Anderson Cancer Center) at the 7th MAQC meeting at SAS Institute: "***If a Statistical Analysis Plan (SAP) is hardwired to a specific data set, we will more likely end up fitting to noise.***"

External Validation Data Sets: Unique to the MAQC-II (thanks to our data providers), we will have at least one confirmatory ("validation") data set for five of the six disease/study areas (Table 1). The timing for the release of these precious data sets is critical and we are doing our best to retain the independency of these data sets from the model development process.

It is my understanding that many data analysis teams are spending a lot of time exploring the distributed MAQC-II training data sets and are expected to have a DAP developed by September 2007. In fact, several teams told me that their data analysis procedures are (almost) fixed and are being applied to different data sets.

Each data analysis team's DAP should be reviewed, commented and agreed upon at least by other data analysis teams in a completely transparent manner, i.e., not anonymous. I am sure that each data analysis team will welcome input from those not directly involved in data analysis but are willing to contribute their expertise to improve the DAP. Thus, all the MAQC-II participants (including those in the RBWG) who are interested in improving the DAPs are encouraged to participate in the open review and comment process. If a data analysis team's DAP is agreed upon by the majority of other data analysis teams, I would suggest that a copy be sent to RBWG for further comments, keeping in mind that all reviews (including those among data analysis teams and by RBWG) are meant to be advisory, and not mandatory, to the data analysis team.

I encourage all of you actively help improve the DAPs through direct interactions with the data analysis teams. I think that this will give your constructive comments and expertise the best chance of enhancing the DAPs. It is important for us to keep in mind that the MAQC-II project will be peer-reviewed as one single team when our manuscripts are submitted for publication. We must work like one team now.

Table 5 of the "MAQC-II Research Plan": Populating the matrix of performance metrics

		Performance Metrics						
		1	2	3	.	.	.	*m*
Analysis Methods (Models)	1	$PM_{1,1}$	$PM_{1,2}$	$PM_{1,3}$	.	.	.	$PM_{1,m}$
	2	$PM_{2,1}$	$PM_{2,2}$	$PM_{2,3}$	.	.	.	$PM_{2,m}$
	3	$PM_{3,1}$	$PM_{3,2}$	$PM_{3,3}$	.	.	.	$PM_{3,m}$
	.							
	.							
	.							
	.							
	n	$PM_{n,1}$	$PM_{n,2}$	$PM_{n,3}$	.	.	.	$PM_{n,m}$

Predictive Modeling Analysis Factors
Russ Wolfinger, May 25, 2007 @ the 7th MAQC face-to-face meeting

1. Initial Preprocessing

None
Background Subtraction
Other instrument-specific methods

2. Data Transformation

None
Log
Shifted Log
Generalized Log
Cube root
Rank

3. Data Summarization (for probe-level data)

Mean
Median
Median Polish
Trimmed Mean
MAS5, PLIER
MBEI, MMEI

4. Data Normalization (on rows of wide matrix)

None
Mean
Median
Mean and Variance
Median and Absolute Deviation
InterquartileRange
Quantileto reference quantiles
Normalization to reference data of some type

5. Basic Predictor Reduction

None
Filtering based on instrument-specific flags
Filtering based on simple statistics, e.g. drop a fraction of low intensity and/or low variance genes

6. Predictor Standardization (on columns of wide matrix)

None
Center to Mean Zero
Scale to Unit Variance
Other Location/Scale standardization

7. Statistical Predictor Reduction

K-Means representatives
Gene-by-gene ANOVA / mixed model (includes t-test and heterogeneous t-test as special case)
Permutation tests
Nonparametric (e.g. Wilcoxon)
Bayesian and EB methods
High breakdown methods

8. Predictive Modeling Methods

Discriminant
Distance Scoring
General Linear Modeling
K Nearest Neighbors
Logistic Regression
Partial Least Squares
Partition Trees
Radial Basis Machine / SVM

9. Cross-Validation Methods

Random Splits
Stratified Random
K-Fold Random
K-Fold Block (includes LOO)
K-Fold Split

10. Performance Criteria

Accuracy
Specificity and Sensitivity
AUC, pAUC
RMSE

How to Cover Millions of Combinations?

Possibility: Experimental Design Approach
Consider previously described aspects as factors in an experimental design.
Use Design of Experiments (DOE) to derive a subset of combinations that cover the space
Assume no higher than two-way interactions between factors
Predictive modeling factor likely the largest because of nesting of options

MAQC2 TOXICOGENOMICS WORKING GROUP TELECONFERENCE

JULY 11, 2007: Dial toll-free 866-653-7616 (US and Canada); toll number (overseas): 210-795-6025; participant passcode 6907138.

We will have two items on the agenda as we resume our teleconferences:

A. DISCUSSION OF PROPOSAL FOR REVIEW PROCESS FOR DATA ANALYSIS PLANS

We would like to discuss a proposal for the review process in MAQC2.

1) REVIEW DAY: A single, one-day meeting (or Webex) in which the analysis teams, as well as any other member of MAQC2 interested in reviewing the data, would have a chance to discuss their review the data analysis plans (DAP) generated by all the other analysis teams. The Review Day will be scheduled approximately 30 days from the release of the last training dataset in order to allow for a broad application of the algorithms across multiple datasets. Generic Data Analysis Plans for each analysis team will be sent to all members of the analysis teams as well as anyone else within MAQC who would like to review the DAPs a week before Review Day. A table summarizing the results from the DAPs across analysis groups would be distributed a few days before this meeting. On Review Day, feedback from reviewers will be presented in the morning meeting. In the afternoon, the review results for each DAP will be collated. Three possible outcomes will be reported:

a. Consensus Approval

b. Consensus Rejection

c. Conditional Approval

DAPs which receive Consensus Approval on Review Day will trigger release of confirmatory datasets for each analysis group submitting these DAPs. DAPs which receive Conditional Approval will be re-submitted within two weeks after updating with proposed changes to review through email and will be approved if a consensus through email is reached that the changes requested have been completed. Rejected DAPs will be re-submitted in a second Review Day to be scheduled 30 days after the first one.

2) GOAL OF REVIEW: The purpose of this review, as has been the goal of the MAQC collaboration since its inception, will be to learn from the work of a matrix of statisticians and scientists and to freely share the lessons we learn from these analyses.

3) OUTCOME OF REVIEW: The outcome of this review will be to assess whether the results from the analysis of the training sets synthesized in the Data Analysis Plans were ready for a challenge with the confirmatory datasets.

4) TIMELINE:

DAY *ACTIVITY*

0 Release of the last training dataset

23 A Generic Data Analysis Plan (DAP) for each analysis team is distributed to all other analysis teams as well as to any other members of MAQC who would like to participate as reviewers for these DAPs.

30 REVIEW DAY

31 Release of confirmatory datasets for each analysis group submitting DAPs receiving consensus approval on REVIEW DAY.

45 DAPs which receive Conditional Approval on REVIEW DAY are re-submitted after updating with proposed changes to review through email.

60 DAPs rejected on REVIEW DAY will be re-submitted in a second REVIEW DAY.

Any additional review of these data will be welcome, since these will contribute to what we will learn about classifiers, but the release of confirmatory datasets will not be contingent on additional review.

B. THE NIEHS DATASET

Discussion of data analysis for the NIEHS dataset.

Thanks, Federico

Data Analysis Teams (DATs) in the MAQC-II

Leming.Shi@fda.hhs.gov Tel: +1-870-543-7387

December 6, 2007

DAT#	Org. Code	DAP File(s)	Organization	Data Analyst	E-mail	Telephone	Reviewed By
1	CBC	CBC DAP - Liang Zhang.doc	CapitalBio Corporation, China	Xin Meng Qinglan Sun Jianping Wu **Liang Zhang** Sheng Zhu	xmeng@capitalbio.com qlsun@capitalbio.com jpwu@capitalbio.com lzhang@capitalbio.com shzhu@capitalbio.com	 +86-13641099545	
2	CAS[1]	CAS DAP - Tieliu Shi I.doc CAS DAP - Tieliu Shi II.doc CAS DAP - Tieliu Shi III.doc	Chinese Academy of Sciences, China	Lei Chen Jian Cui Guang Li **Tieliu Shi** Chen Zhao	skychen21@yahoo.com.cn cuijiananthem@163.com lg0319@126.com tieliushi@yahoo.com zhaochen_ny@yahoo.com.cn	 +86-13918033701	
4	CDRH	CDRH DAP - Gene Pennello.doc	Center for Devices and Radiological Health, FDA	Samir Lababidi Francisco Martinez-Murillo **Gene A. Pennello** Reena Philip Daya Ranamukhaarachchi Rong Tang Zivana Tezak	samir.lababidi@fda.hhs.gov francisco.martinez@fda.hhs.gov gene.pennello@fda.hhs.gov reena.philip@fda.hhs.gov daya.ranamukhaarachchi@fda.hhs.gov rong.tang@fda.hhs.gov zivana.tezak@fda.hhs.gov	240-276-3110 240-276-3943 240-276-3149 240-276-1286 301-796-0238 240-276-3041 240-276-0772	
5	CDRH2	Wagner et al. Plan for MAQC2.doc	Center for Devices and Radiological Health, FDA	Weijie Chen **Robert F. Wagner** Waleed A. Yousef	weijie.chen@fda.hhs.gov robert.wagner@fda.hhs.gov wyousef@aucegypt.edu	301-796-2663 301-796-2529	
6	CIPF	CIPF DAP - Joaquin Dopazo.doc	Centro de Investigacion Principe Felipe, Spain	Fatima Al-Shahrour Ana Conesa **Joaquin Dopazo** Ignacio Medina David Montaner	flshahrour@cipf.es aconesa@cipf.es jdopazo@cipf.es imedina@cipf.es dmontaner@cipf.es	 +34-963289680	
7	Cornell	ICB DAP baseline - Fabien Campagne.doc	Weill Medical College of Cornell University, Institute for Computational Biomedicine (ICB)	**Fabien Campagne** Francesca Demichelis Igor Segota	fac2003@med.cornell.edu frd2004@med.cornell.edu igs2002@med.cornell.edu	646-962-5613 646-962-5642	
8	DKFZ	DKFZ DAP - Benedikt Brors.doc	German Cancer Research Center, Germany	Peter Bewerunge **Benedikt Brors** Lars Kaderali Yvonne Koch Jasmin Mueller Frederik Roels Thomas Wolf Marc Zapatka Matthias Fischer André Oberthuer	p.bewerunge@dkfz-heidelberg.de b.brors@dkfz-heidelberg.de l.kaderali@dkfz-heidelberg.de y.koch@dkfz-heidelberg.de j.mueller@dkfz-heidelberg.de f.roels@dkfz-heidelberg.de t.wolf@dkfz-heidelberg.de m.zapatka@dkfz-heidelberg.de matthias.fischer@uk-koeln.de andre.oberthuer@uk-koeln.de	 +49-6221-42-3602 +49-221-478-6852	
11	EPA	EPA DAP - Richard Judson.doc	U.S. Environmental Protection Agency	Fathi Elloumi **Richard Judson** Zhen Li	elloumi.fathi@epamail.epa.gov judson.richard@epa.gov li.zhen@epamail.epa.gov	919-541-5526 919-541-3085 919-541-4104/966-5421	

13	GeneGO	GeneGo DAP - Weiwei Shi.doc	GeneGo Inc.	Richard Brennan Andrej Bugrim Yuri Nikolsky **Weiwei Shi**	richard@genego.com andrej@genego.com yuri@genego.com helprhelp@gmail.com; weiwei@genego.com	408-454 4050 269-983-7629 858-756-7996 269-983-7620	
14	FBK	FBK-MPBA DAP - Cesare Furlanello.doc	Fondazione Bruno Kessler, Italy	Davide Albanese **Cesare Furlanello** Giuseppe Jurman Stefano Merler Silvano Paoli Samantha Riccadonna	albanese@fbk.eu furlan@fbk.eu jurman@fbk.eu merler@fbk.eu paoli@fbk.eu riccadonna@fbk.eu	+39 0461 314 580 +39 0461 314 523	
16	Ligand	Ligand Pharma DAP - Wen Luo.doc	Ligand Pharmaceuticals Inc.	**Wen Luo**	wluo@ligand.com	858-550-4415	
17	NCTR	NCTR DAP - Weida Tong.doc	National Center for Toxicological Research, FDA	Minjun Chen Xiao-hui Fan Hong Fang Huixiao Hong Roger G. Perkins Leming Shi Zhenqiang Su **Weida Tong** Qian Xie	minjun.chen@fda.hhs.gov xiao-hui.fan@fda.hhs.gov hong.fang@fda.hhs.gov huixiao.hong@fda.hhs.gov roger.perkins@fda.hhs.gov leming.shi@fda.hhs.gov zhenqiang.su@fda.hhs.gov weida.tong@fda.hhs.gov qian.xie@fda.hhs.gov	870-543-7057 870-543-7507 870-543-7538 870-543-7296 870-543-7049 870-543-7387 870-543-7059 870-543-7142 870-543-7219	
18	NIEHS[2]	NIEHS Chou Bushel DAP - Pierre Bushel I.doc NIEHS Li Bushel DAP - Pierre Bushel II.doc	National Institute of Environmental Health Sciences, NIH	**Pierre Bushel** Jeff Chou Jianying Li	bushel@niehs.nih.gov chou1@niehs.nih.gov li11@niehs.nih.gov	919-316-4564 919-316-4529 919-316-4612	
18'	NIEHS2	NIEHS DAP - Jennifer Fostel.doc	National Institute of Environmental Health Sciences, NIH	**Jennifer M. Fostel**	fostel@niehs.nih.gov	919-541-5055	
19	NWU	NWU DAP - Simon Lin.doc	Northwestern University	Pan Du **Simon Lin**	dupan@northwestern.edu s-lin2@northwestern.edu	312-695-1331	
20	Princeton	Princeton DAP - Jianqing Fan.doc	Princeton University	**Jianqing Fan** Yi Ren Yichao Wu Feng Yang	jqfan@Princeton.EDU fyprinceton@gmail.com yichaowu@Princeton.EDU ren@nel-exchange.rutgers.edu	609-258-7924	
21	Roche	Roche DAP - Mark Fielden.doc	Roche Palo Alto LLC	**Mark Fielden** Hans Bitter Andreas Buness Matthew Cooper Guido Steiner Chun Zhang	mark.fielden@roche.com hans.bitter@roche.com bunessa@roche.com matthew.cooper.mc1@roche.com guido.steiner@roche.com zhangc12@roche.com	650-855-5136	
22	SAI	SAI_DAP - John Zhang.pdf	Systems Analytics Inc.	Jun Luo Guohua Ma **John Zhang**	junl@systemsanalytics.com guohuam@systemsanalytics.com johnz@systemsanalytics.com	781-398-2233 x 304 781-398-2233 x 301	
23	SAS	SAS DAP - Russ Wolfinger.doc	SAS Institute Inc.	Wenjun Bao Tzu-Ming Chu Shannon Conners Wendy Czika Mark Lambrecht Stan Martin Padraic Neville Thomas Pedersen Doug Robinson	wenjun.bao@sas.com tzu-ming.chu@sas.com shannon.conners@jmp.com wendy.czika@sas.com mark.lambrecht@sbx.sas.com stan.martin@sas.com padraic.neville@sas.com thomas.petersen@jmp.com Doug.Robinson@jmp.com	919-531-1484 919-531-4675 919-531-2157 +32(0)45960658	

				Susan Shao Pei-Yi Tan **Russell D. Wolfinger**	susan.shao@sas.com pei-yi.tan@sas.com russ.wolfinger@sas.com	 919-531-6695	
24	SIB	SIB DAP - Vlad Popovici.doc	Swiss Institute of Bioinformatics, Switzerland	Mauro Delorenzi Darlene Goldstein **Vlad Popovici**	mauro.delorenzi@isb-sib.ch darlene.goldstein@epfl.ch vlad.popovici@isb-sib.ch	+41-21 692 4097 +41-21 692 4088	
25	Spheromics	Trygg Bylesjo Scherer DAP - Andreas Scherer.doc	Spheromics, Finland	Max Bylesjo **Andreas Scherer** Johan Trygg	max.bylesjo@chem.umu.se Andreas.Scherer@Spheromics.com johan.trygg@chem.umu.se	 +358 400 138 510	
26	SuperArray	SA DAP - Guozhen Liu.doc	SuperArray Bioscience Corporation	**Guozhen (Gordon) Liu** Xiao Zeng	gzliu@superarray.net xzeng@superarray.net	301-682-9200 x 172 301-682-9200 x 108	
27	Tsinghua	Tsinghua University DAP - Xuegong Zhang.doc	Tsinghua University, China	Shicai Fan Bo Jiang Rui Jiang Shao Li Ying Liu Lu Meng Yunfei Pei **Xuegong Zhang**	fsc03@mails.tsinghua.edu.cn bo-jiang02@mails.tsinghua.edu.cn ruijiang@tsinghua.edu.cn shaoli@tsinghua.edu.cn l03@mails.tsinghua.edu.cn casablanca9968@gmail.com peiyf05@mails.tsinghua.edu.cn zhangxg@tsinghua.edu.cn	 +86-10-62794919	
29	Cedars-UCLA	Cedars-Sinai UCLA DAP - Xutao Deng.doc	University of California at Los Angeles Cedars-Sinai Medical Center	**Xutao Deng** Charles Wang Jun Xu	xutao.deng@cshs.org charles.wang@cshs.org jun.xu@cshs.org	310-423-7363 310-423-7361	
30	UIUC	UIUC DAP - SHeng Zhong.doc	University of Illinois at Urbana-Champaign	Jun Cai **Sheng Zhong**	juncai@uiuc.edu szhong@uiuc.edu	 217-265-6589	
31	UML	UML DAP - Dalila Megherbi.doc	University of Massachusetts Lowell	Manual Madera **Dalila B. Megherbi**	manual_madera@student.uml.edu dalila_megherbi@uml.edu	978-934-2889 978-934-3077	
32	USM	USM DAP - Venkata Thodima.doc	University of Southern Mississippi	Youping Deng Ying Li **Venkata Thodima**	youping.deng@usm.edu ying.li@usm.edu venkata.thodima@usm.edu	601-266-6678	
35	Almac	Almac DAP - Juergen Frese.doc	Almac Diagnostics, UK	**Juergen von Frese**	juergen.frese@almacgroup.com	+44-28 3839 5978	
36	ZJU	ZJU DAP - Xiaohui Fan.doc	Zhejiang University, China	**Xiaohui Fan** Yiyu Cheng Yong Mao Haibin Qu Yi Wang	fanxh@zju.edu.cn xiaohuifan@gmail.com chengyy@zju.edu.cn ymao@iipc.zju.edu.cn quhb@zju.edu.cn mysky@zju.edu.cn	+86-571-87951138	
37	JHSPH	JHSPH DAP - Rafael Irizarry.doc	Johns Hopkins Bloomberg School of Public Health	**Rafael Irizarry** Matthew McCall	ririzarr@jhsph.edu	410-614-5157	

[1]Three DAPs were submitted from the same Data Analysis Team (DAT) from the CAS; [2]Two DAPs were submitted from the same NIEHS DAT. 'There are two independent DATs from the NIEHS.

The Matthews Correlation Coefficient (MCC) as a Possible Performance Metric for Assessing the Quality of Classifiers Being Developed by the MAQC Project

Huixiao.Hong@fda.hhs.gov & Leming.Shi@fda.hhs.gov

September 13, 2007

We suggest that the **Matthews Correlation Coefficient** (MCC) be considered as a metric for all data sets for assessing the performance of binary classification models being developed by the MAQC consortium. We offer this suggestion in hope of stimulating on-going discussions within MAQC RBWG on performance metrics. Comments on our proposal and alternative proposals from MAQC participants are most welcome.

The MCC is defined as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}$$

It returns a value between -1 and +1. A coefficient of +1 corresponds to totally correct predictions, 0 to an average random prediction, and -1 to totally incorrect predictions. A positive MCC value corresponds to better than random prediction (informative) and a negative value to a worse than random (uninformative) prediction. Importantly, the MCC similarly characterizes prediction accuracy regardless of the distribution of samples between classes. In contrast, a global metric such as *Accuracy* can be misleading when samples are unbalanced between classes.

While there is no perfect way of characterizing the confusion matrix of true and false positives and negatives with a single number, the MCC is generally regarded as being one of the best such measures. The MCC uses all four numbers (*TP*, *TN*, *FP*, and *FN*) and often provides a more balanced evaluation of the model than other commonly used performance metrics such as:

Accuracy = (TP + TN) / TS
Sensitivity = TP / AP = TP / (TP + FN)
Specificity = TN / AN = TN / (TN + FP)
Positive Predictive Value (PPV) = TP / PP = TP / (TP + FP)
Negative Predictive Value (NPV) = TN / PN = TN / (TN + FN)

The Confusion Matrix

Total Samples (*TS*)	Actual Positives (*AP*)	Actual Negatives (*AN*)
Predicted Positives (*PP*)	True Positives (*TP*)	False Positives (*FP*)
Predicted Negatives (*PN*)	False Negatives (*FN*)	True Negatives (*TN*)

References:

1. Matthews BW (1975). Comparison of the prediction and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**: 442-451.
2. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**: 412-424.
3. http://en.wikipedia.org/wiki/Matthews_Correlation_Coefficient

Summary of the WebEx Meeting on MAQC-II Review of Data Analysis Protocols (DAPs)

WebEx Meeting Date: January 9, 2008
Summary Date: January 13, 2008

Leming.Shi@fda.hhs.gov (Tel: +1-870-543-7387)

The objectives of this meeting were:

- To discuss the review comments on the 33 Data Analysis Protocols (DAPs) developed by 28 Data Analysis Teams (DATs);
- To ensure that MAQC-II participants understand the next steps and timelines for the project;
- To decide on a few important issues related to data analysis and submission of results.

Over 50 participants attended the WebEx, representing the vast majority of the 28 data analysis teams. Each data analysis team attending the WebEx summarized the comments made on its DAP. Review comments were made available to data analysis teams at the MAQC-II ftp site. The meeting Agenda and associated important information on MAQC-II data analysis can be found in "MAQC-II_Review_DAPs_Agenda_Attachments_09JAN2008.pdf", which was distributed by Leming Shi before the WebEx.

DAP submission and review processes:

Tim Davison (Asuragen) and Lakshmi Vishnuvajjala (FDA/CDRH) described the DAP submission and review processes coordinated by the Regulatory Biostatistics Working Group (RBWG). The extended deadline for DAP submission was November 13, 2007. Each submitted DAP was made available to all data analysis teams and assigned to four reviewers for comments, two from the "neighboring" data analysis teams in numerical order and the other two from the RBWG volunteers.

Summary of the 33 DAPs:

Thirty-three (33) DAPs were submitted to the MAQC-II from 28 data analysis teams (organizations), with contributors from at least nine countries (US, China, Finland, Germany, Italy, Spain, Sweden, Switzerland, and UK). The MAQC-II consortium is grateful to the data analysis teams for developing their thoughtful DAPs and for the individuals who provided constructive comments on the DAPs for further improvement. The 33 DAPs explored a large number of combinations of various modeling factors, including:

- Normalization methods: 14 normalizations methods have been used in at least one DAP, with MAS5, RMA, and gcRMA being the most frequently chosen methods for the Affymetrix platform data;
- Classification methods/algorithms: 50 classification methods have been used in at least one DAP, with SVM, KNN, and random forest being the most frequently chosen methods;
- Similarly, there were many options for each of other modeling steps such as feature filtering, feature selection, and batch-effect removal;
- Internal validation: Six internal validation procedures were proposed to estimate a model's performance and/or to guide the model development process, with 10-fold cross-validation

(10-CV), leave-one-out cross-validation (LOOCV), and 5-fold cross-validation (5-CV) being the mostly frequently chosen approaches;

- Performance metrics: 12 performance metrics have been proposed in at least one DAP to measure the overall performance of a model and its prediction of confirmatory data, with overall accuracy, AUC (Area Under the Receiver Operating Characteristic Curve), and the Matthews Correlation Coefficient (MCC) being the most frequently used.

MAQC-II is not a competition:

Leming Shi emphasized that the MAQC-II project was not meant to be a competition among participants to develop the "best" model for each of the six data set (13 endpoints); instead, it is a collaborative research project aims to develop a general data analysis procedure that is likely applicable to future data sets. Therefore, the purpose of the DAP is not to produce the "best" model for a particular data set, but rather to evaluate many data analysis workflows and to determine whether a workflow performs reasonably well on different data sets (endpoints); that is why we are working on multiple data sets (endpoints). It is expected that some workflows will be more robust (work on most data sets), whereas other workflows may be more prone to overfitting and may not extrapolate well to external validation data sets. An important goal of the MAQC-II is to determine the relative importance of modeling factors to a model's performance so that a general data analysis procedure that is likely to work outside of the MAQC-II data sets may be identified and recommended as MAQC-II's "best" practices. Each of the many models to be developed by the MAQC-II will be characterized by several performance metrics (Y), for both internal and external validation, and a sequence of decisions (e.g. normalization and classification methods) to be made along the modeling steps (X). Mining this "matrix of performance matrix" is expected help us identify the most important factors for developing predictive models; therefore, "good" and "bad" models are equally important to reach this goal and should be submitted. Each model represents a sample in the modeling space.

Performance metrics:

We had lengthy discussions on the choice of metric(s) for measuring the performance of a model and its prediction of confirmatory (validation/blind) data sets. With understanding that there is no single "best" metric for all situations, we decided that the following five metrics be calculated for each model for performance evaluation:

- Matthews Correlation Coefficient (MCC)
- Overall Accuracy (ACC)
- Sensitivity (SEN)
- Specificity (SPE)
- AUC (Area Under the Receiver Operating Characteristic Curve)

These five metrics should be considered as the minimum set of performance metrics and must be calculated for each model to be submitted to the MAQC-II by the DAP teams. Although each DAP team has the freedom to calculate additional performance metrics of its preference, the aforementioned five metrics must be provided for each submitted model. Upon receiving the prediction results from each DAP team on each confirmatory data set (endpoint), the MAQC-II will calculate these five metrics to judge a model's performance in external validation. Leming Shi expressed his strong preference of using MCC as the primary metric for measuring a model's performance in external validation. Each DAP team should exercise its own best judgment in