Replicability of predictions across studies: challenges and opportunities

giovanni_parmigiani@dfci.harvard.edu

Statistical Methods and Models for Complex Data,
Padova, September 2022

disclosures

- Consulting Related to the Topic:
  Martingale Labs, Delfi Diagnostics
- Speaker's Bureau: None
- Grant/Research support from: NIH-NCI, NSF
  Licensing of BayesMendel software for genetic counseling.
  Licensing of Ask2me database.
- Stockholder in: Phaeno Biotechnology
- Honoraria from: Academic Only
- co-Founder / Chief Scientific Officer: Phaeno Biotechnology
- Relevant patents: POSTN for debulking in OC
- Patents on diagnostic use of various genes
- Employee of: Dana Farber Cancer Institute

none of the analyses described involve licensed products

reproducibility and replication: a crisis?

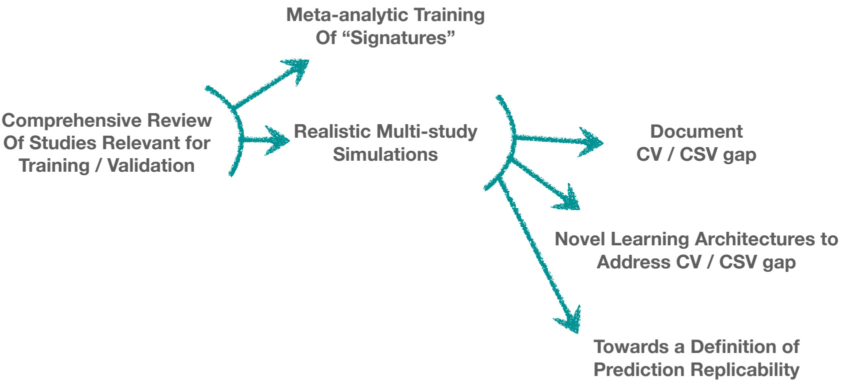
An **ad hoc committee of the National Academies** of Sciences, Engineering, and Medicine explored the issues of reproducibility and replication in scientific and engineering research, focusing on defining reproducibility and replicability, and examining the extent of non-reproducibility and non-replicability.

reproducibility and replication in prediction

Word counts:

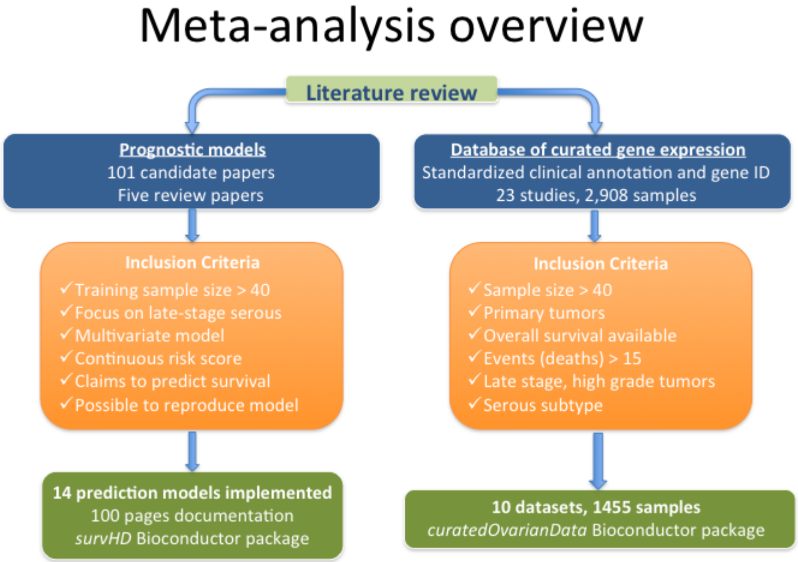
hypothesis test	**27**
machine learning	**3**

outline and flowchart



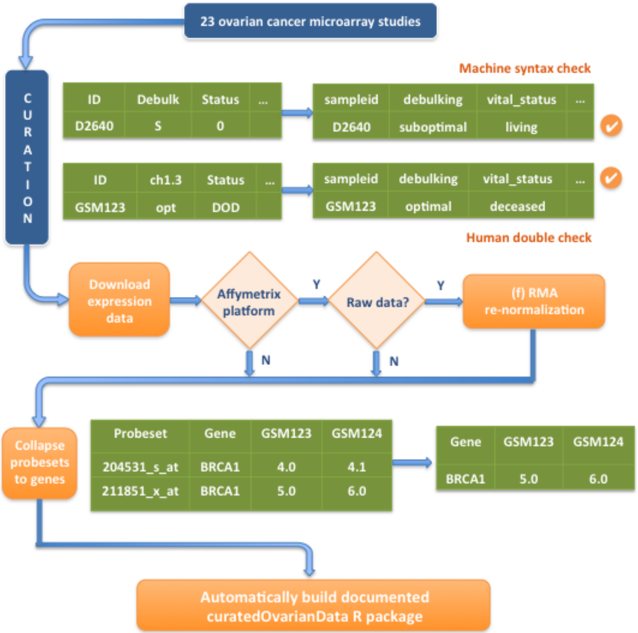
classifier validation via meta-analysis

Waldron etal JNCI 2014



CuratedOvarianData

Ganzfried etal Databases 2012

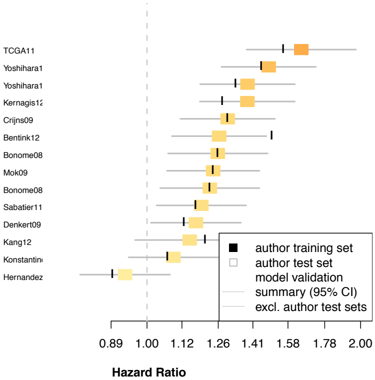


classifier cross-study validation

Waldron etal JNCI 2014

Implemented Models — Validation Statistics for 14 Models in 10 Datasets

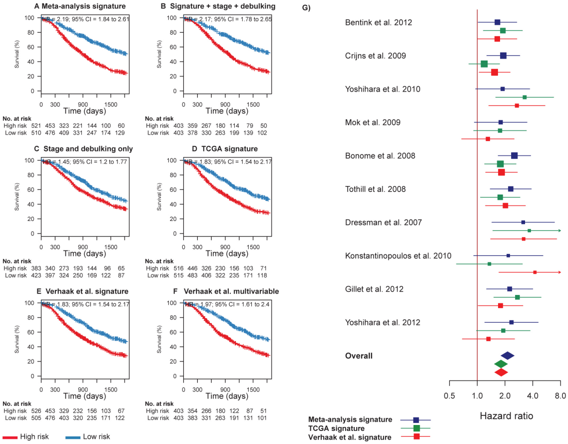
Implemented Models / **Expression Datasets**	Dressman	Yoshihara 2012A	Mok	Tothill	Konstantinopoulos	Bonome	Bentink	TCGA	Crijns	Yoshihara 2010
Dataset Average	1.81	1.47	1.43	1.41	1.39	1.37	1.35	1.14	1.11	1.04
TCGA11	2.05	2.28	1.58	1.85	1.36	1.64	1.97	1.94	1.07	1.53
Yoshihara12	2.44	9.65	1.21	1.8	1.02	1.77	1.97	1.21	1.35	1.42
Yoshihara10	2.69	1.38	1.15	1.93	1.45	1.57	1.47	1.33	0.7	7.27
Kernagis12	2.65	1.39	2.91	2.08	1.45	1.39	1.25	1.23	1.32	0.87
Crijns09	1.22	1.92	1.49	1.51	1.2	1.44	1.28	1.1	3.04	1.21
Bentink12	1.94	1.01	1.89	1.44	1.2	1.14	1.62	1.16	1.26	1.45
Bonome08_263genes	1.3	2.73	2	1.32	0.53	2.01	1.45	1.17	1.03	0.77
Mok09	1.54	1.82	3.18	1.71	0.89	1.58	1.28	0.98	0.95	1.39
Bonome08_572genes	0.8	1.89	1.1	1.41	2.29	2.27	1.47	1.07	1.35	0.84
Sabatier11	1.95	1.15	1.17	1.41	1.72	1.07	1.19	1.11	1.3	0.73
Denkert09	2.6	0.76	1.33	1.31	2.25	1.04	1.29	1.08	1.15	0.79
Kang12	2.14	1.19	0.81	0.85	1.21	1.46	1.17	1.55	1.02	0.73
Konstantinopoulos10	1.34	1.01	0.82	1.07	2.05	1.07	1	1.15	0.97	1.09
Hernandez10	0.68	0.55	1.07	0.71	0.86	1.21	0.79	1.04	0.9	1.03



training a classifier by meta-analysis

Riester etal JNCI 2014

Figure 4. Combined comparison of our novel meta-analysis gene signature with existing prognostic factors and signatures ...



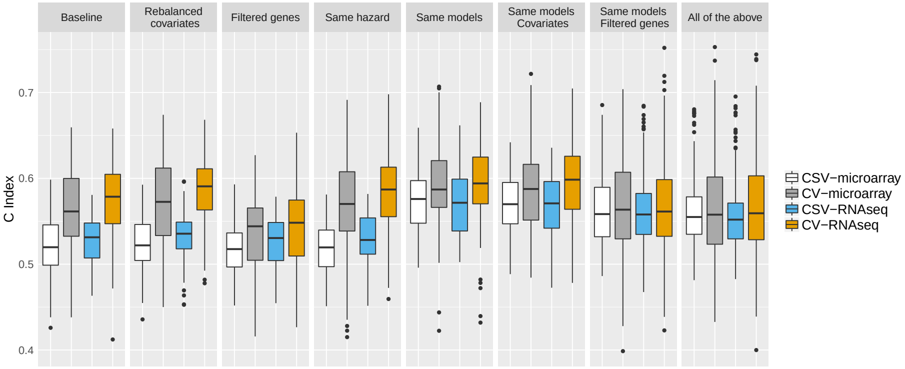
J Natl Cancer Inst, Volume 106, Issue 5, May 2014, dju048, https://doi.org/10.1093/jnci/dju048

Generates collections of studies

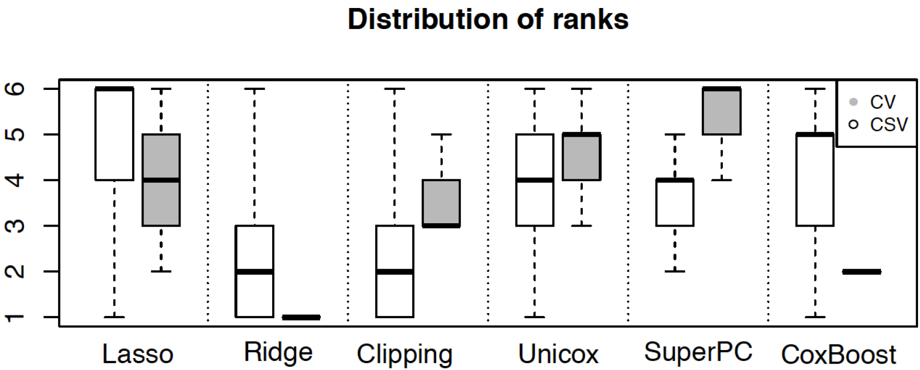
Within and across study variation is closely matching empirical collections based on comprehensive reviews.

determinants of CV/CSV gap

Zhang etal Biostatistics 2018



CV and CSV rank methods differently



Attention to replicability
requires rethinking existing machine learning principles.

How do we engineer statistical learning methods to validate well out of sample?

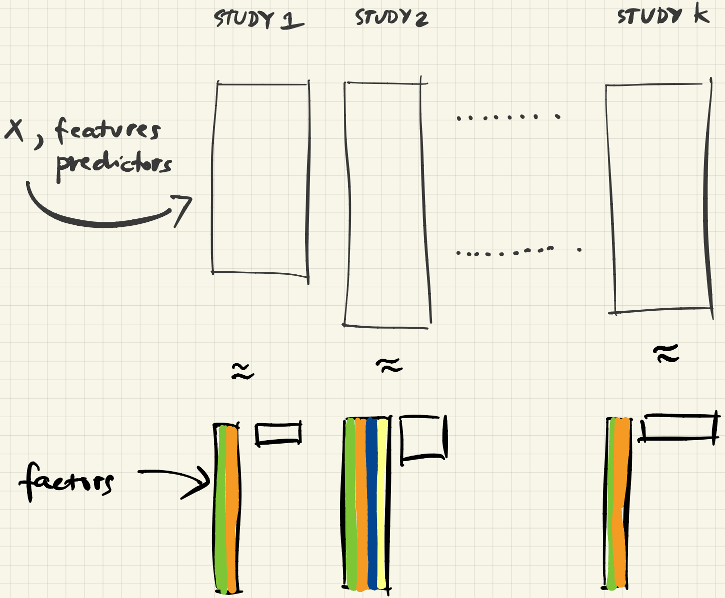
Use multiple studies for training.

Keywords	Meta-analysis, Domain Adaptation
Niche	Simple, Scalable and Interpretable Architectures at the Statistical Learning / Health Interface

Unsupervised multi-study learning

multi-study factor analysis

DeVito Biometrics 2018 and AAS 2019



standard factor analysis on two studies

DeVito 2018

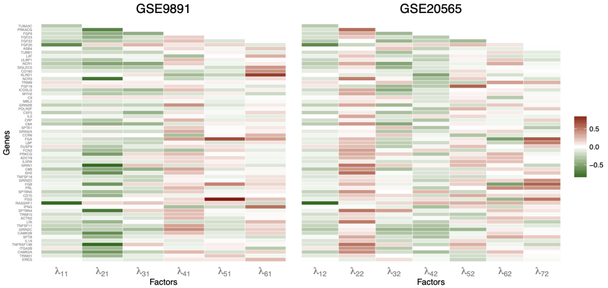
a.

Table 1

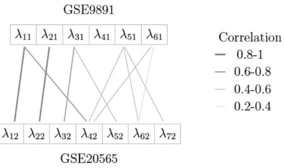
The four data sets considered in the illustration and their characteristics.

Study	Samples	Platform	Late Stage (%)	Reference	T_s
GSE9891	285	Affy U133Plus 2.0	85	Tothill et al. (2008)	6
GSE20565	140	Affy U133Plus 2.0	48	Meyniel et al. (2010)	7
GSE26712	195	Affy U133a	96	Bonome et al. (2008)	10
TCGA	578	Affy HT U133a	90	Cancer Genome Atlas Research Network (2011)	9

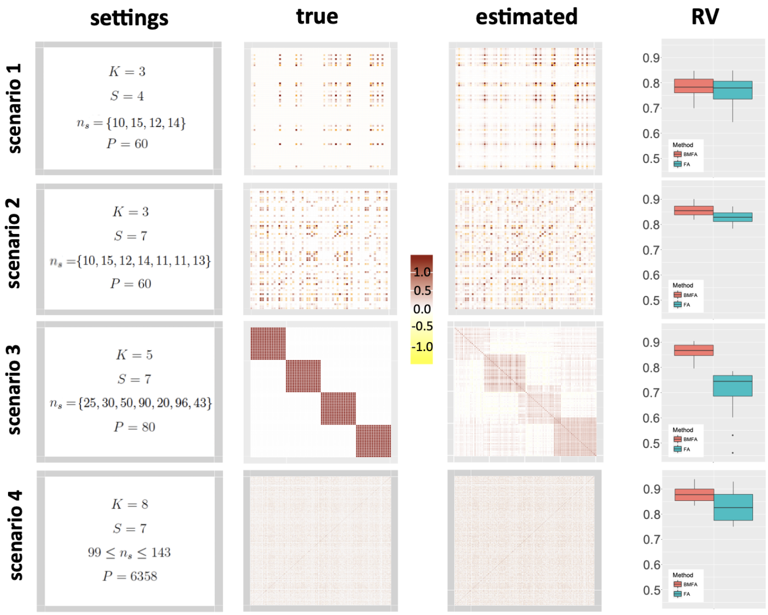
b.



c.

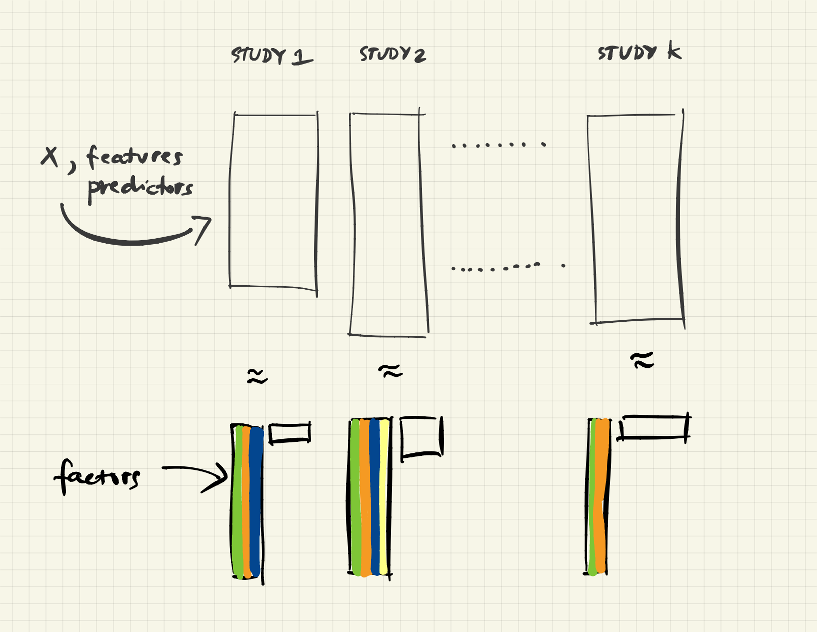


recovery of covariance matrices



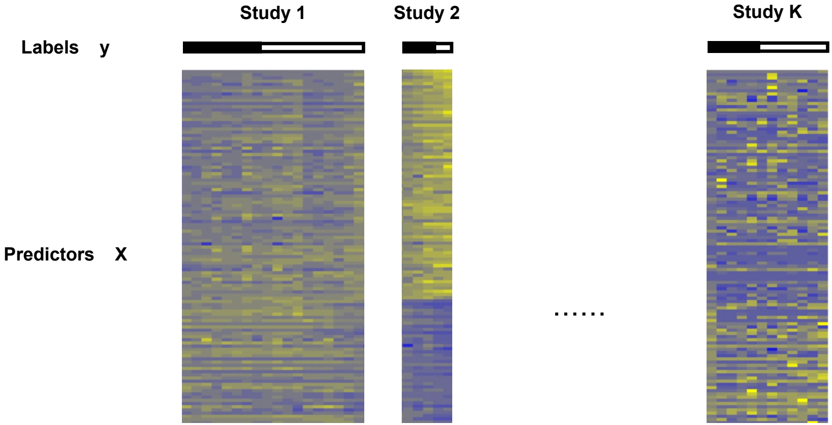
multi-study factor analysis: combinatorial

Grabski AAS in press

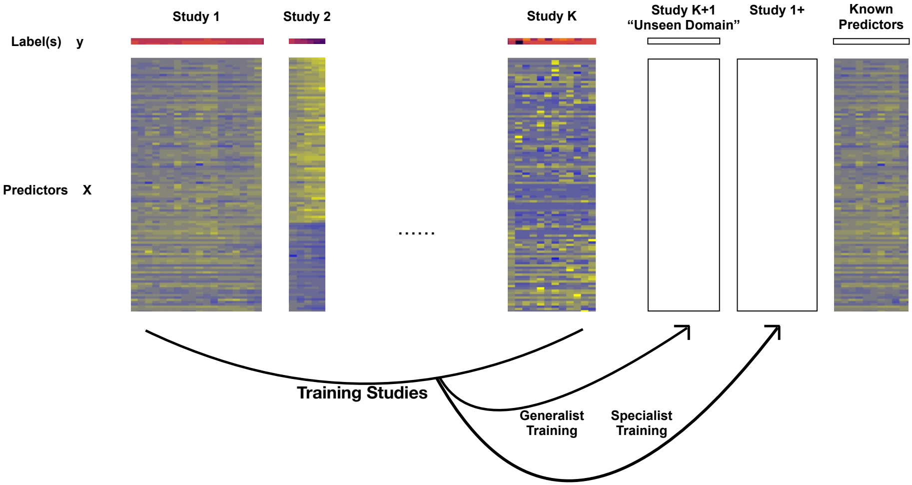


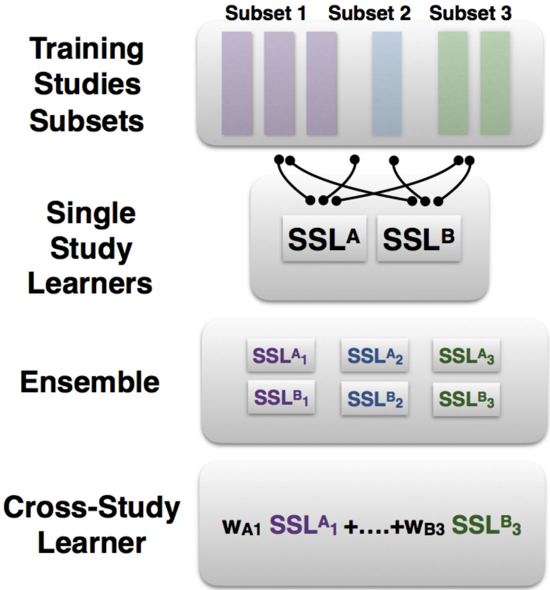
Supervised multi-study learning

supervised data structure



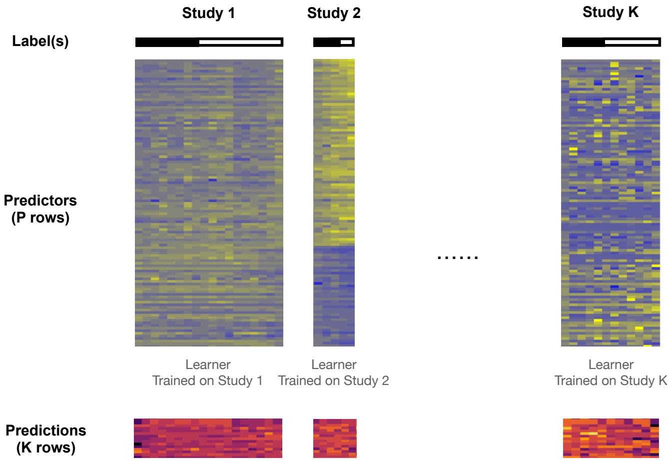
multi-study learning: goals




Training Studies Subsets
Subset 1
Subset 2
Subset 3
Single Study Learners
SSLA
SSLB
Ensemble
SSLA1
SSLA2
SSLA3
SSLB1
SSLB2
SSLB3
Cross-Study Learner
WA1 SSLA1 +....+WB3 SSLB3

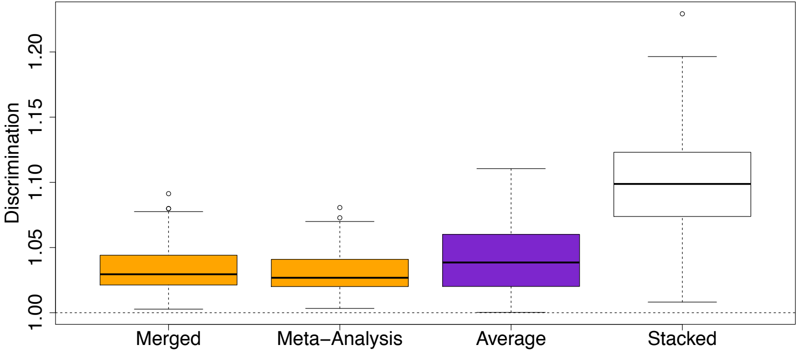
"generalist" multi-study stacking

Patil & Parmigiani PNAS 2018



Stage I	Separately	train learners to predict y_k on X_k by study
Stage II	Jointly	train a learner to predict y on T

ovarian cancer studies



TRAINING STUDIES COMBINED INTO ONE

MODEL COEFFICIENTS META-ANALYZED

MULTI-STUDY EQUAL WEIGHT

MULTI-STUDY NEAR-OPTIMAL WEIGHTS

from here

	"transition point"
Guan 2019	for random coefficients generating model
Shyr 2022	for boosting
Ramchandran 2019	ensembling forests vs trees
Ramchandran 2021	cross-cluster weighted forests
Loewinger 2019	"study strap"
	a continuum between merging and MSS
Ren 2020	multi-study stacking as optimization
	no-data-reuse training
	asymptotics
Loewinger 2021	optimal ensemble construction

Towards a definition of replicability

decision theoretic definition

Characters (1,2, or 3):

Modeler, prediction or scoring rule ϕ

Agent, decision problem

Assessor(s), with gold standard studies $S_1, \ldots, S_K$

Replicability: Assessor(s) agree that the modeler's tool,
in the context of a specific decision problem,
is providing similar average utility across studies.

Modeler+Agent

The Modeler+Agent holds:

prediction or scoring rule ϕ

model π on $\mathcal{X}$ and $\mathcal{Y}$

utility $U(a, y) : (\mathcal{A} \times \mathcal{Y}) \to \mathbb{R}$.

An optimal decision function δ^* satisfies

$$\delta^*(\phi(x)) = max_{\delta \in \Delta} E_\pi \left\{ U(\delta(\phi(x)), y) \right\}$$

Assessor(s)

The prediction rule ϕ is replicable if its optimal application to the same decision problem in different data sets leads to approximately the same average utility to the decision maker. Formally:

Definition (Absolute ϵ-replicability)

ϕ is ϵ-replicable in absolute utility over $S_1, \ldots, S_K$ if

$$\max_{k,k'} |\mathcal{U}_k - \mathcal{U}_{k'}| \leq \epsilon$$

where, for study k, the agent's utility is, on average,

$$\mathcal{U}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} U(\delta^*(\phi(x_{ik})), y_{ik}) \tag{1}$$

example: Classification Replicability

A classification algorithm $\varphi : \mathcal{X} \to \mathcal{Y}$
e.g. $\varphi = \delta^*(\phi(x)) = max_{\delta \in \Delta} E_\pi \left\{ U(\delta(\phi(x)), y) \right\}.$

Utility function defined directly as

$$U(\varphi(x), y) : (\mathcal{Y} \times \mathcal{Y}) \to \mathbb{R}$$

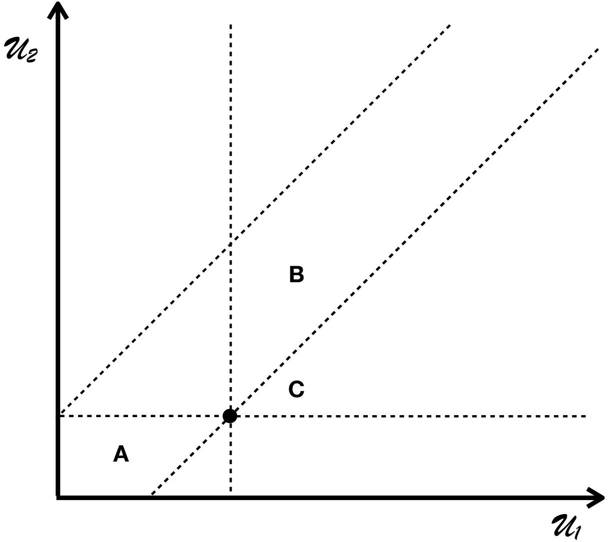
$\mathcal{U}_k$ defined as

$$\mathcal{U}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} U(\varphi(x_{ik})), y_{ik}) \tag{2}$$

and apply Definition 1.

e.g if $U(\varphi, y) = I_{\varphi = y}$ then $\mathcal{U}_k$ is the empirical correct classification proportion in study k
and ϵ-replicability obtains when this proportion does not vary by more than ϵ in any two-study comparison.

Dominance



take home messages

"Science of data science"
There is value in building and analyzing
collections of related studies
to understand real world properties of statistical methods.

"Multi-study Learning"
There is value in using multiple studies
to improve built-in replicability.

Much remains to be investigated
from a theoretical point of view.

credits and references: cross-study validation

Ganzfried, B.F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X.V., Ahmadifar, M., Birrer, M.J., Parmigiani, G., Huttenhower, C., Waldron, L., 2013. curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. Database 2013, bat013–bat013.

Waldron, L., Haibe-Kains, B., Culhane, A.C., Riester, M., Ding, J., Wang, X.V., Ahmadifar, M., Tyekucheva, S., Bernau, C., Risch, T., Ganzfried, B.F., Huttenhower, C., Birrer, M., Parmigiani, G., 2014. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. JNCI 106, dju049–dju049.

Zhao, S.D., Parmigiani, G., Huttenhower, C., Waldron, L., 2014. Más-o-menos: a simple sign averaging method for discrimination in genomic data analysis. Bioinformatics 30, 3062–3069.

Bernau, C., Riester, M., Boulesteix, A.-L., Parmigiani, G., Huttenhower, C., Waldron, L., Trippa, L., 2014. Cross-study validation for the assessment of prediction algorithms. Bioinformatics 30, i105–12.

Trippa, L., Waldron, L., Huttenhower, C., Parmigiani, G., 2015. Bayesian nonparametric cross-study validation of prediction methods. Ann. Appl. Stat 9, 402–428.

Waldron, L., Riester, M., Ramos, M., Parmigiani, G., Birrer, M., 2016. The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. JNCI 108, djw146.

Zhang Y, Bernau C, Parmigiani G, Waldron L. The impact of different sources of heterogeneity on loss of accuracy from genomic prediction models. Biostatistics. 2018;6(e1002358):701.

credits and references: multi-study learning

Patil, P., Parmigiani, G. Training replicable predictors in multiple studies. Proc Natl Acad Sci USA. 2018;115(11):2578-2583. doi:10.1073/pnas.1708283115.

Guan, Z., Parmigiani, G., Patil, P. Merging versus Ensembling in Multi-Study Machine Learning: Theoretical Insight from Random Effects. arXiv:1905.07382

Ramchandran M, Patil P, Parmigiani G. Tree-weighting for multi-study ensemble learners. Pacific Symposium on Biocomputing 451-462 2020.

Loewinger G, Acosta Nunez R, Mazumder R, Parmigiani G. Optimal Ensemble Construction for Multi-Study Prediction with Applications to COVID-19 Excess Mortality Estimation. arXiv, https://arxiv.org/abs/2109.09164.

Ramchandran M, Mukherjee R, Parmigiani G. Cross-Cluster Weighted Forests. 2021 arXiv.2105.07610

Shyr C, Sur P, Parmigiani G, Patil P. Multi-Study Boosting: Theoretical Considerations for Merging versus Ensembling. `https://arxiv.org/pdf/2207.04588.pdf`

Zhang Y, Patil P, Johnson WE, Parmigiani G. Robustifying Genomic Classifiers To Batch Effects Via Ensemble Learning. Bioinformatics 11 2020. Btaa986.