Statistical methods for learning to classify with biomarker panels: insights from cryptanalysis

Paul McKeigue, Marco Colombo

Usher Institute of Population Health Sciences and Informatics

Methodological problems of learning to predict from biomarker panels

- What sample size is required to learn to classify from a high-dimensional biomarker panel?
- How should predictive performance of the panel be reported?
- How do we learn a model that:-
  - has optimal predictive performance
  - is explainable (not "black-box")
  - uses the smallest subset of biomarkers

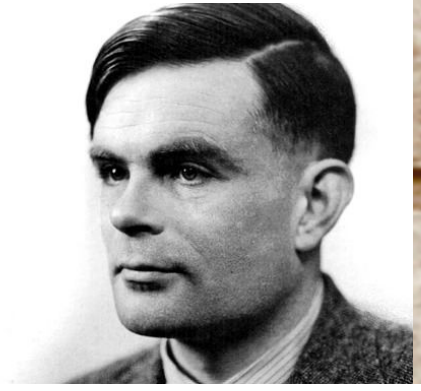
Limitations of the C-statistic (area under ROC curve) for assessing predictive performance

- Not obvious how C is relevant to risk stratification
- Increment in C obtained by adding new biomarkers is difficult to interpret
    - depends on what covariates were included in the baseline model, and whether they were matched between cases and controls
- Mistaken belief that no useful increase in predictive performance can be achieved by adding new biomarkers to a baseline model that has $C > 0.9$
- Widely-adopted alternatives - "Integrated Discrimination Improvement" and "Net Reclassification Index" - are not (mathematically) proper scoring rules.

A Bayesian approach to assessing predictive performance

- *Weight of evidence* $W\left(\frac{\text{case}}{\text{control}}\right)$ favouring case over control status provided by a score is the logarithm of the Bayes factor (ratio of likelihoods).
  - can calculate it on test data for any predictor that outputs probabilities
  - weights of evidence contributed by independent predictors are additive
- Sampling distribution of $W\left(\frac{\text{case}}{\text{control}}\right)$ in cases and controls defines how the score will perform as a risk stratifier
  - if only we knew this sampling distribution

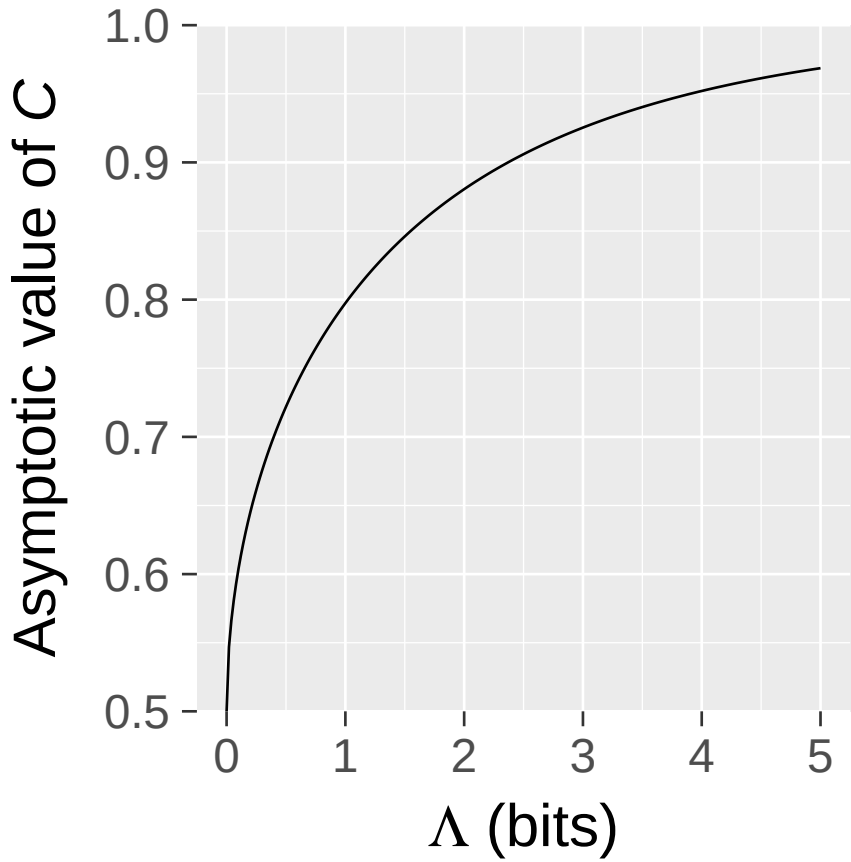
Hut 8, Bletchley Park 1941


If the effective number of independent predictors is large:-

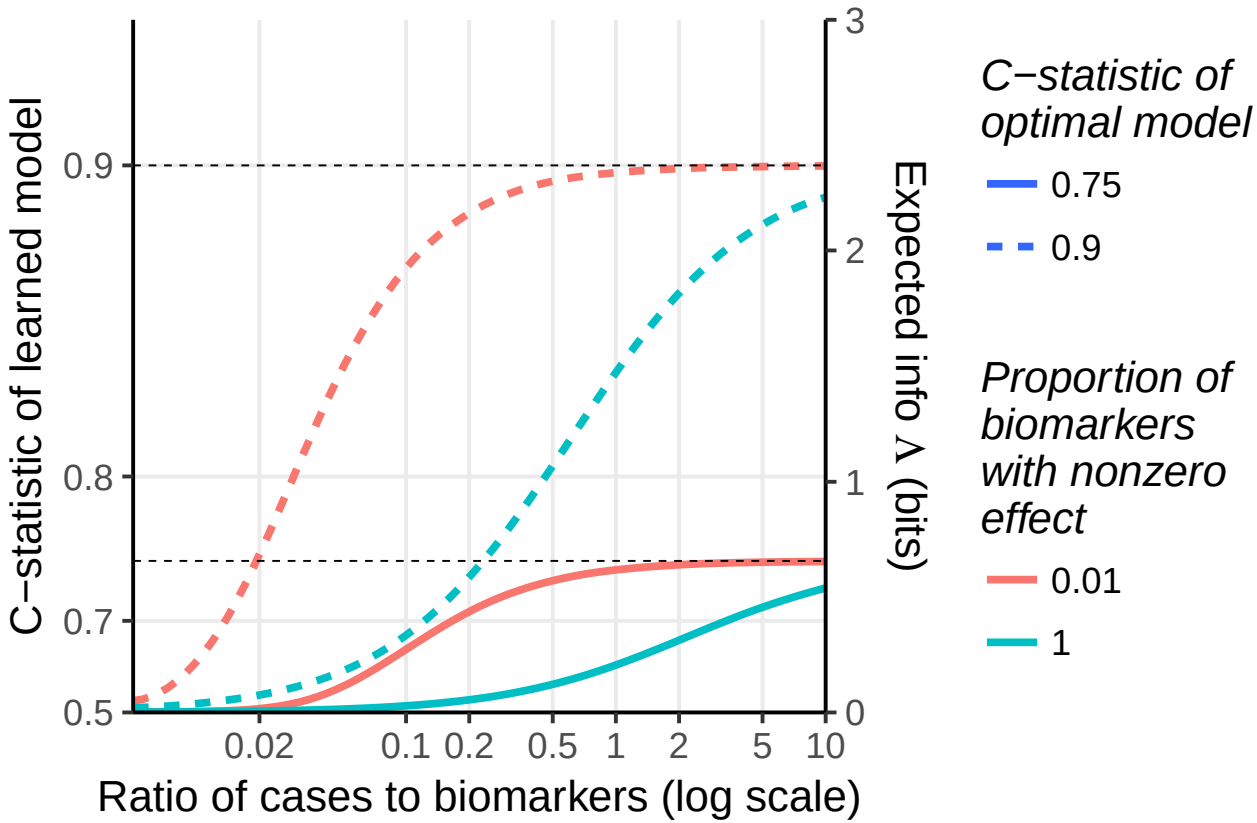
- distributions of $W\left(\frac{\text{case}}{\text{control}}\right)$ in cases and $W\left(\frac{\text{control}}{\text{case}}\right)$ in controls are gaussian with expectation Λ, variance 2Λ natural log units
  - we only need Λ to characterize how the score will perform as a risk stratifier
- Λ is the *expected information for discrimination*

Relation of C-statistic to expected information for discrimination Λ



- Contributions of independent predictors are additive on scale of Λ, but not on scale of C.

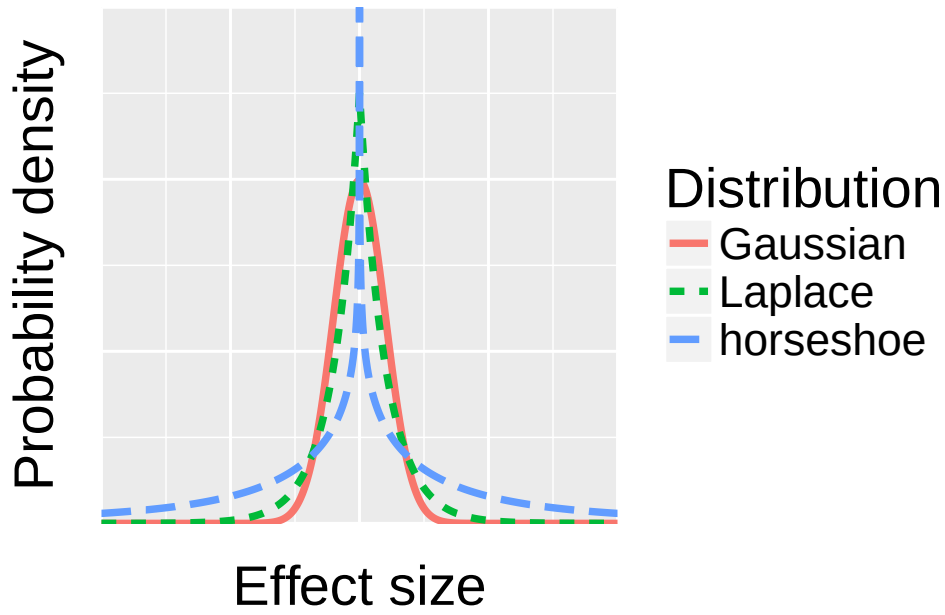
Learning curves for classifying with high-dimensional biomarker panels (McKeigue, *SMMR* 2017)



- Required ratio of cases to biomarkers to learn 80% of information for discrimination:
  - 0.1 for $C = 0.9$, 1% of biomarkers with nonzero effect
  - 9 for $C = 0.75$, all biomarkers with nonzero effect

Models for prediction from high-dimensional biomarker panels

- To model biomarker effects, we have to model the distribution of effect sizes
  - many biomarkers of tiny effect, or a few biomarkers of large effect?
  - standard penalized regression methods - ridge, LASSO - are not flexible enough



Stan (Carpenter, Gelman, Hoffman et al 2017) - a program for Bayesian computation

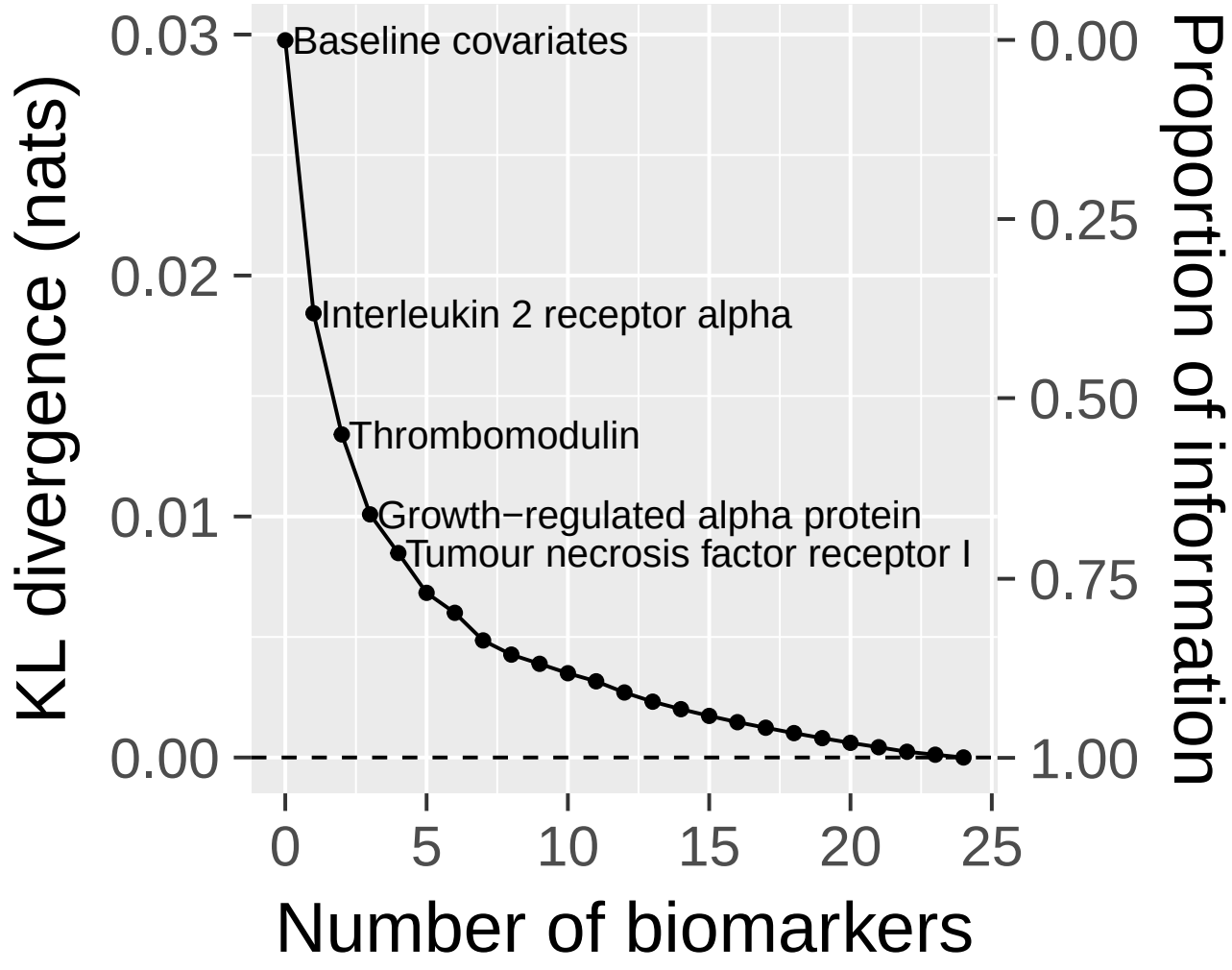
- Bayesian theory always works - but in practice the computation may be intractable
  - Standard Bayesian learning algorithms cannot learn models with many correlated parameters
- *Stan* overcomes this by using *Hamiltonian Monte Carlo* algorithm (Duane, Kennedy, Pendleton & Roweth 1987)
  - samples the posterior distribution of all parameters jointly
  - Can learn model of biomarker effects, including distribution of effect sizes, in one step

Evaluating a panel of biomarkers for prediction of rapid progression of diabetic nephropathy

- 789 individuals from Scottish Type 1 Bioresource
- 24 proteins measured in serum (Myriad/RBM platform).
- Bayesian logistic regression with hierarchical shrinkage distribution of effect sizes, fitted with *Stan*, evaluated by cross-validation.

	C	Λ (bits)	W var/mean (nats)	$\Delta \log \mathcal{L}$ (bits)
Baseline	0.48	0.0		0.0
Baseline + biomarkers	0.60	0.2	2.0	20.6

Projective variable selection of biomarkers, using posterior samples from full model



Conclusions

- Expected information for discrimination Λ should replace C as summary measure of predictive performance
    - estimated by cross-validation, or on a separate test dataset
- Evidence favouring one model over another should be evaluated as difference in test log-likelihoods of models.
- Using *Stan* we can evaluate performance of a biomarker panel and select the best subset of biomarkers in one step.
- Even good tests ($\Lambda > 3$ bits) will often give wrong results