#AACRMolEpi

The Prostate Cancer Clinical Trials Consortium

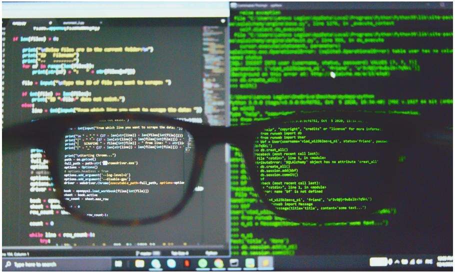
Predictive modeling from high-throughput results

AACR IME Workshop
July 30, 2021

Travis Gerke, ScD
Director of Data Science, PCCTC
gerket@mskcc.org
@travisgerke

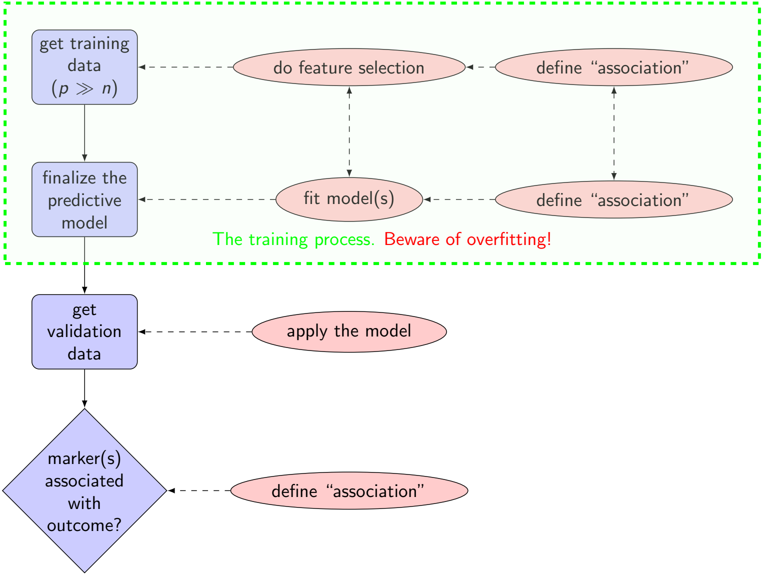
What we're going to talk about

1. Feature selection
2. How do we know models are good?
3. The dangers of overfitting



One view of a high-throughput study

You measure lots of features (p) in some patients (n) with follow-up


get training data ($p \gg n$)
do feature selection
define "association"
finalize the predictive model
fit model(s)
define "association"
The training process. Beware of overfitting!
get validation data
apply the model
marker(s) associated with outcome?
define "association"

Feature selection

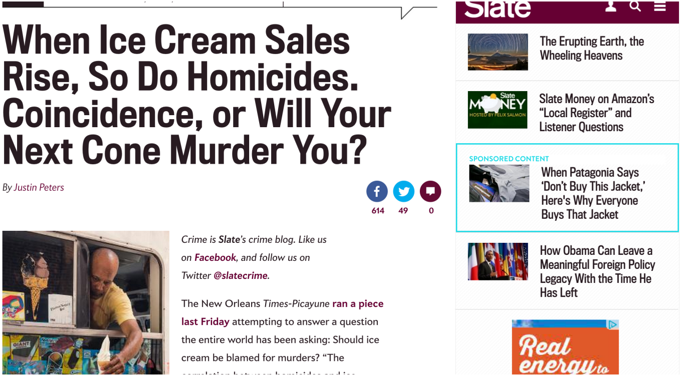
Typical scenario: You've measured thousands of genes/SNPs/markers and need to know which ones are important

- You probably also have some clinical variables too

Your criteria for including variables in your model depends on the goal

- Causal determinants/intervention targets: choose with expert input
- Risk prediction (prognostication): choose whatever works

So, for prediction, fear not the confounded relationship!



When Ice Cream Sales Rise, So Do Homicides. Coincidence, or Will Your Next Cone Murder You?

By Justin Peters

*Crime is **Slate**'s crime blog. Like us on **Facebook**, and follow us on Twitter **@slatecrime**.*

The New Orleans *Times-Picayune* **ran a piece last Friday** attempting to answer a question the entire world has been asking: Should ice cream be blamed for murders? "The

The Erupting Earth, the Wheeling Heavens

Slate Money on Amazon's "Local Register" and Listener Questions

How Obama Can Leave a Meaningful Foreign Policy Legacy With the Time He Has Left

Feature selection: abbreviated list of methods

Exhaustive search: pick best model from all variable combinations

- Becomes computationally awful; 30 variables $\Longrightarrow$ 1.1 trillion models

Greedy search: in a stepwise fashion, pick the next best variable(s)

- Forward and backward selection are examples of this method
- Faster, but with the tradeoff that you might not find the best set

Penalized regression such as lasso or elastic net

- Convenient to think of these as fancy greedy algorithms
- Fast, easy to interpret

Variable importance measures from ensemble methods

- Example: random forests consist of many classification trees, where each tree "votes" for the outcome. One can assess which variables were critical for accuracy in many trees.

Information-based filtering (only unsupervised method on this page)

- Examples: Remove all features correlated at > 0.70; PCA filtering
- This does not select based on how well outcome is predicted!

Heuristic: choose your own adventure!

- Make your own rules. If it validates in independent data, you win.

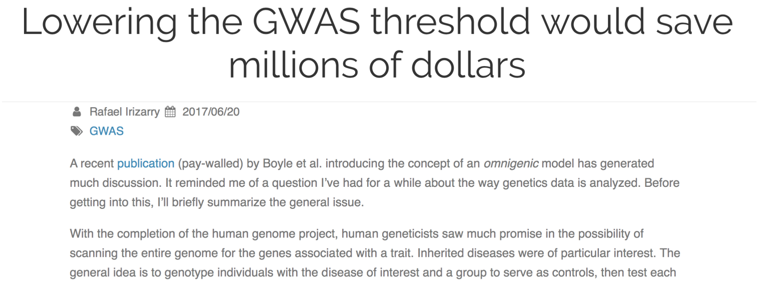
Feature selection: inference in training

Note that *p*-value thresholds were not mentioned on the previous slide

- But this is not the way GWAS studies work right? (think 10^{-8})
- Interestingly, methods for expression studies have evolved differently, particularly in that there is no standard cutoff
- But both types of studies have the goal of independent validation

Exercise: You're doing the first ever GWAS. When will you use *p*-values (training, validation, neither), and what cutoffs will you use?

- Recent controversial motivation below [link]



Lowering the GWAS threshold would save millions of dollars

Rafael Irizarry 2017/06/20

GWAS

A recent publication (pay-walled) by Boyle et al. introducing the concept of an *omnigenic* model has generated much discussion. It reminded me of a question I've had for a while about the way genetics data is analyzed. Before getting into this, I'll briefly summarize the general issue.

With the completion of the human genome project, human geneticists saw much promise in the possibility of scanning the entire genome for the genes associated with a trait. Inherited diseases were of particular interest. The general idea is to genotype individuals with the disease of interest and a group to serve as controls, then test each

Feature selection: opinionated list of tools

- Awesome: R/Bioconductor
  - Suggested starting packages: `caret`/`parsnip`, `glmnet`, `pamr`, `RWeka`
  - For deep learning: `Keras`
  - Stepwise selection available through the `step() function`
  - An easy language in which to write your own algorithms
  - Massive benefit to having bioinformatics packages for iterative QC
- Awesome: Python, often through `scikit-learn`
  - Can be a bit harder to install/maintain than R/RStudio
  - Note: `scikit-learn` is amazing for ML, but may provide unexpected statistical inference in some specific settings [link]
- Less awesome: Stata is somewhere in the middle
  - Some modules exist: `CHAIDFOREST`, `LARS`, and `ELTMLE`
  - Interestingly, some of those modules rely on R ¯_(ツ)_/¯
- Less awesome: SAS is more limiting, but it can be done
  - Standard SAS can do forward/backward selection in standard models
  - Need Enterprise Miner for random forest, penalized regression, PCA

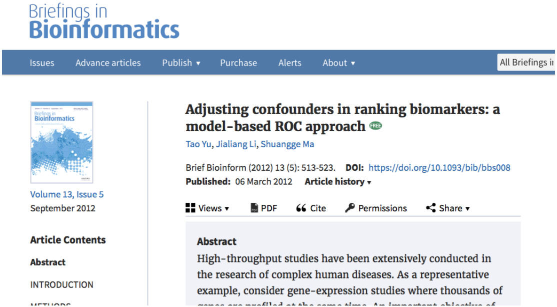
Feature selection: what to do with clinicopathologic factors?

Assuming you're not the lead biostatistician on the project, a cursory familiarity with the methods/tools slides will likely suffice!

How you treat known predictive factors in your study will, however, be critical; and this is not a solely statistical concern

Exercise: tumor grade is highly predictive of eventual PrCa metastasis. You profiled tumor tissue to find a gene expression signature to predict metastasis. What do you do with the variable "grade"?

Bonus: what are "confounders" in predictive models?



Volume 13, Issue 5
September 2012

Article Contents

Abstract

INTRODUCTION

METHODS

Adjusting confounders in ranking biomarkers: a model-based ROC approach

Tao Yu, Jialiang Li, Shuangge Ma

Brief Bioinform (2012) 13 (5): 513-523. **DOI:** https://doi.org/10.1093/bib/bbs008
Published: 06 March 2012 **Article history** ▾

Abstract

High-throughput studies have been extensively conducted in the research of complex human diseases. As a representative example, consider gene-expression studies where thousands of genes are profiled at the same time. An important objective of

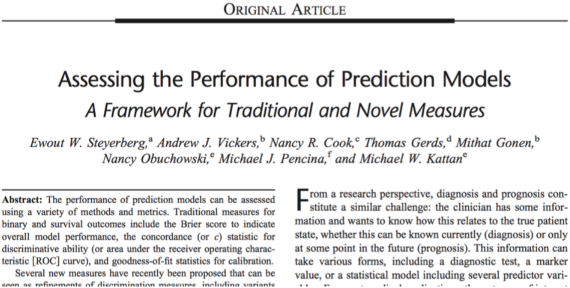
Model fit: How do we measure "good'?

For binary markers, we have some standard metrics

- Sensitivity: $se = P(M^+ \mid D^+)$; specificity: $sp = P(M^- \mid D^-)$
- Positive predictive value: $PPV = P(D^+ \mid M^+)$
- Negative predictive value: $NPV = P(D^- \mid M^-)$

But risk exists on a continuum $\implies$ probabilistic classifiers

- Evaluating prediction accuracy more difficult when we can't fit results into a 2×2 table



ORIGINAL ARTICLE

Assessing the Performance of Prediction Models

A Framework for Traditional and Novel Measures

Ewout W. Steyerberg,[a] Andrew J. Vickers,[b] Nancy R. Cook,[c] Thomas Gerds,[d] Mithat Gonen,[b] Nancy Obuchowski,[e] Michael J. Pencina,[f] and Michael W. Kattan[e]

Abstract: The performance of prediction models can be assessed using a variety of methods and metrics. Traditional measures for binary and survival outcomes include the Brier score to indicate overall model performance, the concordance (or *c*) statistic for discriminative ability (or area under the receiver operating characteristic [ROC] curve), and goodness-of-fit statistics for calibration.

Several new measures have recently been proposed that can be seen as refinements of discrimination measures, including variants

From a research perspective, diagnosis and prognosis constitute a similar challenge: the clinician has some information and wants to know how this relates to the true patient state, whether this can be known currently (diagnosis) or only at some point in the future (prognosis). This information can take various forms, including a diagnostic test, a marker value, or a statistical model including several predictor variables. For most medical applications, the outcome of interest

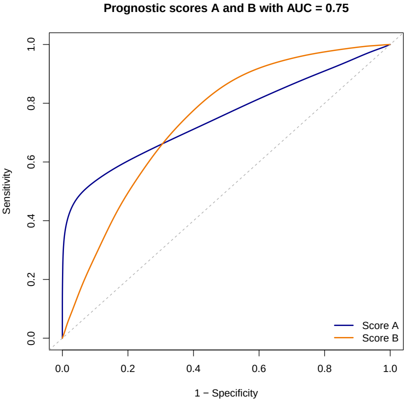
Model fit: Discrimination

How well the model separates cases from non-cases

- Distinct from prediction accuracy: could predict all cases with $\hat{p} = 0.51$ and non-cases with $\hat{p} = 0.50$ for perfect discrimination

The most common statistic: area under the ROC curve (AUC)

- The probability that a randomly selected case will have a higher predicted risk than a randomly selected non-case



Model fit: Measures related to clinical utility

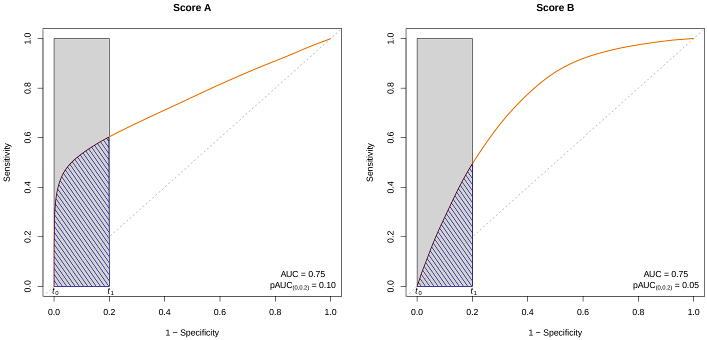
Most of the above measures assume the “cost” of a false negative is the same as that of a false positive

In practice, this is most often not the case!

Reclassification: how does a new marker shift predicted classifications?

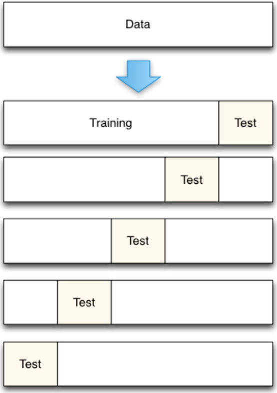
Decision curve analysis: visualize/test the benefit of competing models

Partial area under the ROC curve (pAUC): only calculate AUC for areas clinically applicable



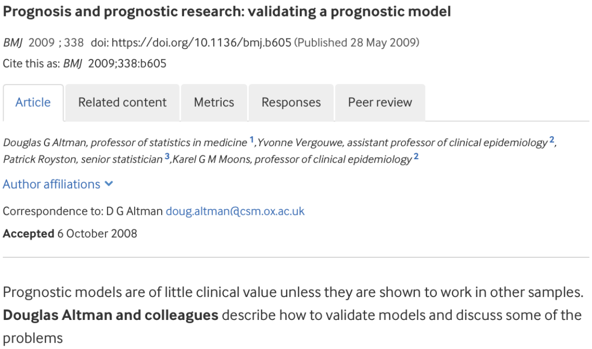
The dangers of overfitting: cross-validation

- When you have $p > n$, you can *always* find a perfect model
    - But this model is extremely unlikely to work in other data
    - If you do this, you have overfit your data
- One strategy to guard against overfitting: cross-validation
    - If possible, nest *everything* you do in the training data in the cross-validation step



The dangers of overfitting: external validation

- Even if you use cross-validation, you may have still overfit and selected noisy features of your training data
  - External validation in an independent data source is *critical*!
  - If this is impossible (e.g. due to cost), you can sometimes hold out a portion of your initial data until the very end, though this is not as powerful



Prognosis and prognostic research: validating a prognostic model

BMJ 2009 ; 338 doi: https://doi.org/10.1136/bmj.b605 (Published 28 May 2009)

Cite this as: *BMJ* 2009;338:b605

Douglas G Altman, professor of statistics in medicine [1], *Yvonne Vergouwe, assistant professor of clinical epidemiology* [2], *Patrick Royston, senior statistician* [3], *Karel G M Moons, professor of clinical epidemiology* [2]

Correspondence to: D G Altman doug.altman@csm.ox.ac.uk

Accepted 6 October 2008

Prognostic models are of little clinical value unless they are shown to work in other samples. **Douglas Altman and colleagues** describe how to validate models and discuss some of the problems

The dangers of overfitting: model complexity

- There is generally a “sweet spot” for model complexity with respect to ultimate performance in test (external) data
  - Cross-validation can help find this sweet spot, but is not a panacea
  - Often the best case is small number of features with strong signal

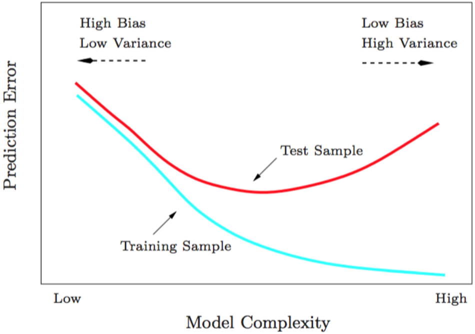


FIGURE 2.11. *Test and training error as a function of model complexity.*

* Figure from Hastie et al. *Elements of Statistical Learning*

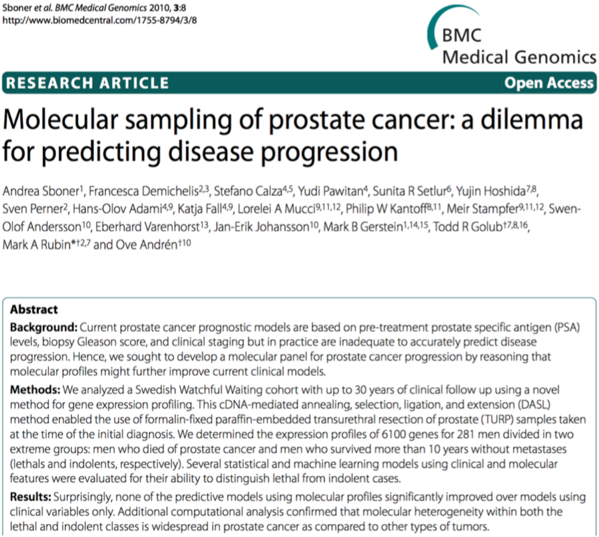
Exercise

You have measured expression of 6000 genes from prostate tumors at diagnosis and have 30 years of follow-up data. You would like to build a prediction model for risk of future metastasis. Which of the following strategies would you use?

- Established clinical parameters (e.g. Gleason grade and stage) alone
- Clinical parameters + top 10 genes from logistic regression
- Clinical parameters + machine learning algorithm using all genes

Important to consider: which is most likely to validate in independent data, and which has the highest risk of overfitting?

Exercise



Sboner *et al. BMC Medical Genomics* 2010, **3**:8
http://www.biomedcentral.com/1755-8794/3/8

BMC
Medical Genomics

RESEARCH ARTICLE **Open Access**

Molecular sampling of prostate cancer: a dilemma for predicting disease progression

Andrea Sboner[1], Francesca Demichelis[2,3], Stefano Calza[4,5], Yudi Pawitan[4], Sunita R Setlur[6], Yujin Hoshida[7,8], Sven Perner[2], Hans-Olov Adami[4,9], Katja Fall[4,9], Lorelei A Mucci[9,11,12], Philip W Kantoff[8,11], Meir Stampfer[9,11,12], Swen-Olof Andersson[10], Eberhard Varenhorst[13], Jan-Erik Johansson[10], Mark B Gerstein[1,14,15], Todd R Golub[†7,8,16], Mark A Rubin[*†2,7] and Ove Andrén[†10]

Abstract

Background: Current prostate cancer prognostic models are based on pre-treatment prostate specific antigen (PSA) levels, biopsy Gleason score, and clinical staging but in practice are inadequate to accurately predict disease progression. Hence, we sought to develop a molecular panel for prostate cancer progression by reasoning that molecular profiles might further improve current clinical models.

Methods: We analyzed a Swedish Watchful Waiting cohort with up to 30 years of clinical follow up using a novel method for gene expression profiling. This cDNA-mediated annealing, selection, ligation, and extension (DASL) method enabled the use of formalin-fixed paraffin-embedded transurethral resection of prostate (TURP) samples taken at the time of the initial diagnosis. We determined the expression profiles of 6100 genes for 281 men divided in two extreme groups: men who died of prostate cancer and men who survived more than 10 years without metastases (lethals and indolents, respectively). Several statistical and machine learning models using clinical and molecular features were evaluated for their ability to distinguish lethal from indolent cases.

Results: Surprisingly, none of the predictive models using molecular profiles significantly improved over models using clinical variables only. Additional computational analysis confirmed that molecular heterogeneity within both the lethal and indolent classes is widespread in prostate cancer as compared to other types of tumors.