Methodology and misunderstandings in precision medicine

Paul McKeigue

Usher Institute of Population Health Sciences and Informatics

Acknowledgements

- Marco Colombo: biomarker predictions, GENOSCORES platform
- Athina Spiliopolou: genetic prediction
- Joe Mellor: deep learning on retinal images
- Andreas Ochs: modelling retinopathy screening
- Helen Colhoun: Scottish Diabetes Research Network Type 1 Bioresource

A new era of precision medicine?

- 2015: US government announced research initiative on "***precision medicine***, *an innovative approach to disease prevention and treatment that takes into account individual differences in people's genes, environments, and lifestyles*".
  - Synonyms: personalized medicine, stratified medicine
- Mathur 2017: "*Precision medicine is an innovative approach towards delivering improved healthcare and* ***reducing overall healthcare costs***"

Methodological problems in precision medicine

1. Learning to subtype disease
   - objective is to learn subtypes of disease that have different outcomes
2. Quantifying predictive performance
   - a better alternative to the C-statistic (area under ROC curve)
3. Is it feasible to predict drug response from biomarkers?
   - success with infectious agents and tumours has not been paralleled in other diseases
   - results in general are poor
4. Regulation and impact on health services
   - does evaluation need randomized trials?
   - will precision medicine reduce health care costs?

1. Learning to subtype disease

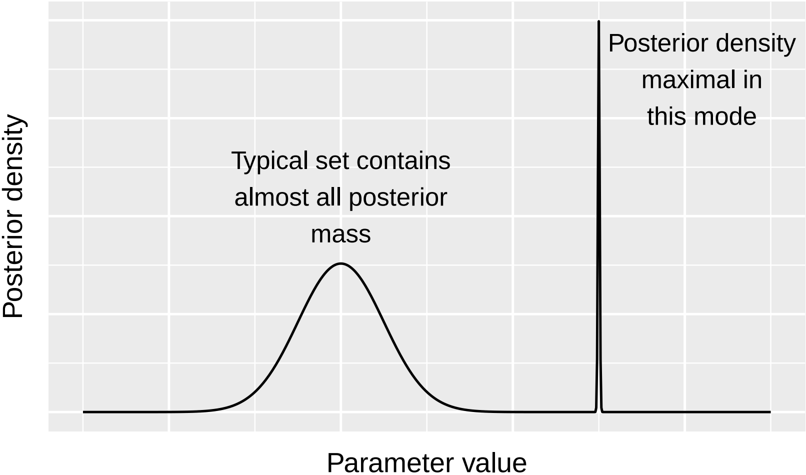
- MRC (2013) workshop on stratified medicine favoured "identifying groups of patients with distinct **endotypes**:"*subtypes of a condition defined by a distinct functional or pathobiological mechanism*"
  - Endotypes should predict response not just to drugs in use now, but to new therapies not yet developed,
  - Soft classification: "patient may traverse more than one endotype during the course of their disease"

Learning to subtype disease: a mixture modelling problem

- Soft classification should allow each individual to be a mix of different endotypes
- Different models for prediction of outcome from covariates may be fitted to each endotype:
- Such models have been described in different contexts
    - in social science as *latent class models* (Lazarsfeld 1968)
    - in biostatistics as *finite mixture models* (Everitt 1981)
    - in machine learning as *mixtures of experts* (Jordan 1994].
- Learning mixture models from data is notoriously difficult

Why learning mixture models from data is hard

- Likelihood surface is multimodal: maximizing the likelihood or posterior density may find an atypical mode
- Model comparison, based on comparing likelihoods of models, is computationally hard



Stan: a platform for Bayesan inference and imputation: Gelman, Lee and Guo (2015)

- Bayesian inference is based on generating the posterior distribution of model parameters given the data
- BUGS and JAGS sample the posterior, updating variables one at a time
- *Stan* uses *Hamiltonian Monte Carlo* (Duane, Kennedy, Pendleton & Roweth 1987) - to update all parameters jointly.
  - momentum as a randomized auxiliary variable
  - algorithmic differentiation to compute gradients
- As an alternative learning algorithm, Stan can use a faster variational Bayes approximation to the posterior.
  - also generates a lower bound approximation to the likelihood of the model (ELBO).

Type 1 diabetes as an exemplar of a disease with underlying endotypes

- Type 1 diabetes is now recognized to be a heterogeneous condition:
  - classic juvenile-onset Type 1 with rapid autoimmune destruction of islet cells
  - late-onset cases in whom loss of beta-cell function progresses slowly, some of whom have features of Type 2 diabetes including obesity
- Residual insulin secretion (measured as C-peptide) may persist years after diagnosis even in early-onset cases.

Scottish Diabetes Research Network Type 1 Bioresource

- Cohort of people clinically diagnosed as Type 1 diabetes over wide ranges of age at onset and duration.
- 5998 individuals with median duration of diabetes 20 years at enrolment.
- C-peptide measured at clinic visit, autoantibodies measured in half the cohort
- genotyped with Illumina chip, untyped SNPs imputed from UK10K reference panel

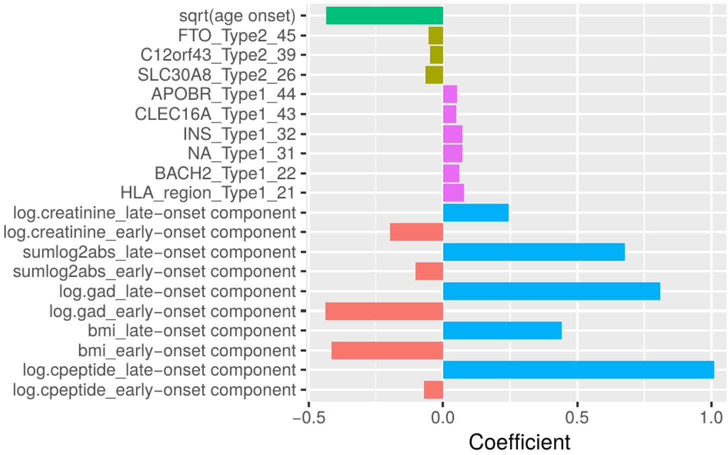
Calculation of genotypic risk scores from GWAS summary statistics

- Genotypic risk scores for Type 1 diabetes and Type 2 diabetes computed using the GENOSCORES platform.
    - genotype vector $\boldsymbol{g}$, genotype correlations $\boldsymbol{\Sigma}$ (estimated from reference panel), univariate regression coefficients $\boldsymbol{\alpha}$ from publicly available summary statistics
    - genotypic risk score is computed as $\boldsymbol{g}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}$
    - Coefficients approximate the weights that would be obtained by fitting a multivariate regression model to the individual-level data.
    - Score computed for each diabetes-associated region

Statistical model

- logistic regression of each individual's mixture component on covariates $\boldsymbol{Z}$: age at onset, genotypic scores for Type 1 and Type 2 diabetes.
  - $\mathrm{logit}\left(\boldsymbol{\lambda}\right) = \boldsymbol{Z}^{\intercal}\gamma$
- linear regressions of J outcome variables on covariates $\boldsymbol{X}$ given kth mixture component:
  - $\langle \boldsymbol{y}_j \mid k \rangle = \boldsymbol{X}^{\intercal}\beta_{jk}$
- y_{ij}: jth outcome variable in ith individual distributed as mixture of component-specific distributions with mixture weights $\lambda_i, (1 - \lambda_i)$

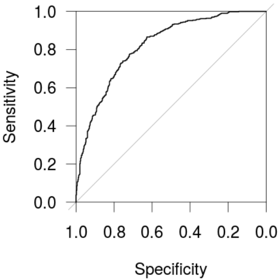
Coefficients (posterior means) of a 2-component mixture model for Type 1 diabetes



Limitations of current version of *Stan* for learning mixture models

- With current version of *Stan* it is possible to learn a model of diabetes as a mix of two endotypes - classic early onset Type 1, late-onset with Type 2-like features
- Neither variational Bayes nor Hamiltonian Monte Carlo explore the multimodal posterior adequately in a single run
    - hack is to use multiple runs of variational bayes algorithm to select best mode, then run Hamiltonian sampler
- New features in the Stan development pipeline may improve learning of multimodal posteriors:-
    - annealing with adiabatic cooling (heat bath)
    - Riemannian Monte Carlo

2. Quantifying predictive performance: an alternative to the area under ROC curve (C-statistic)



- C = probability of correctly classifying a case-control pair
- Proper scoring rule - rewards honest prediction
- Does not require calibration
- Does not depend on prevalence of disease
  - if no covariates in model, C in case-control study is same as in cohort

Problems with the C-statistic

- No obvious application to risk stratification
- Increment in C obtained by adding new biomarkers has no obvious interpretation
    - depends on what covariates were included in the baseline model and whether they were matched
- Only small increments in C can be achieved by adding new biomarkers to a baseline model that has $C > 0.9$
    - mistaken belief that no useful increment in predictive performance can be obtained.

"Researchers have observed that ΔAUC depends on the performance of the underlying clinical model. For example, good clinical models are harder to improve on, even with markers that have shown strong association."

Alternatives to the C-statistic

- Pencina 2008: “Integrated discrimination improvement” and “net reclassification index”
- Hilden and Gerds (2014) - these indices are not proper scoring rules
  - performance can be “improved” by cheating
- Collins 2015:

  “Identifying suitable measures for quantifying the incremental value of adding a predictor to an existing prediction model remains an active research area”.

Bayesian approach to hypothesis testing and classification

Odds form of Bayes theorem (Wrinch and Jeffreys 1921):-

$$(\text{prior odds } \mathrm{H}_1/\mathrm{H}_2) \times \frac{\text{likelihood of } \mathrm{H}_1}{\text{likelihood of } \mathrm{H}_2} = (\text{posterior odds } \mathrm{H}_1/\mathrm{H}_2)$$

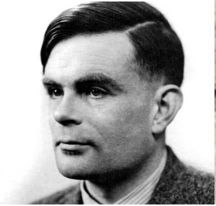
All evidence for inference from data is contained in the likelihood ratio (Bayes factor)
Taking logarithms, this becomes

$$\log(\text{prior odds}) + \text{weight of evidence } \mathrm{H}_1/\mathrm{H}_2 = \log(\text{posterior odds})$$

- Weights of evidence contributed by independent predictors are additive

- Sampling distributions of weight of evidence in cases and controls determine how predictor will behave as risk stratifier

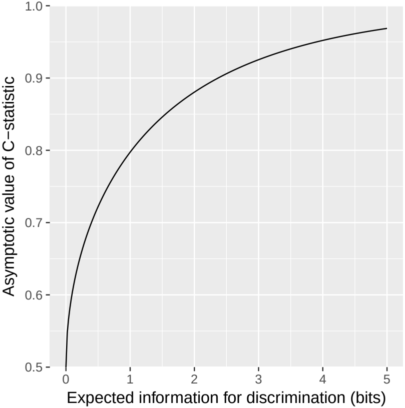
Hut 8, Bletchley Park 1941



If the effective number of independent predictors is large:-

- sampling distribution of the weight of evidence in favour of a true hypothesis is Gaussian
- expectation Λ of weight of evidence in favour of a hypothesis when it is false is minus 1 times its expectation when the hypothesis is true
- variance of weight of evidence is twice its expectation (when natural logarithms are used)

Asymptotic relation of C-statistic to expected log Bayes factor Λ



- Increment of one bit in Λ is asymptotically equivalent to increase in C-statistic from 0.5 (0 bits) to 0.8, or from 0.88 (2 bits) to 0.925.

Using Λ to evaluate predictive performance

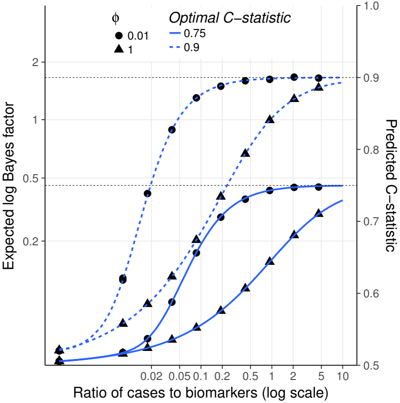
- To Bayesian statisticians: Λ is expected weight of evidence favouring correct assignment
- To informaticians: Λ is *expected information for discrimination* between cases and controls
- Contributions of independent predictors are additive on the scale of Λ.
    - Incremental contribution of a biomarker does not depend on whether cases and controls were matched for covariates
- If Turing's approximation holds, Λ contains all the information needed to characterize how the predictor will stratify risk in a setting with given prior odds of disease.

What value of Λ corresponds to useful prediction?

- Suggested criteria in a clinical setting:-
  - Moderate performance: 1 bit (C=0.80)
  - Good performance: 3 bits (C=0.925)
- For population screening:
  - Moderate performance: 3 bits (e.g. FIT testing for colorectal cancer)
  - Good performance $>$ 5 bits
- Even a good test will often give wrong answers:
  - with $\Lambda = 4$ bits, log-likelihood ratio will be in wrong direction in 12% of individuals tested

Sample size required to learn to classify from high-dimensional biomarker panels

- depends on
  - information content of optimal predictor
  - sparseness of distribution of effect sizes



Genetic prediction: what predictive performance can we expect?

- Polygenic model - many loci of small effect - for genetic effects on disease risk implies additivity on logistic scale
  - principle of maximum entropy
- Clayton (2009): under polygenic logistic model
  - $\Lambda = \log \lambda_S$
  - where λ_S is the sibling recurrence risk ratio
- so if sibling recurrence risk ratio is 2, the optimal info for discrimination is 1 bit (C=0.80).
- with 500K tag SNPs of which 1% are associated with disease risk, sample size required to learn a predictor that extracts 80% of info for discrimination is 50,000
  - larger sample size required if $> 1\%$ of SNPs are associated

Example: genotypic prediction of colorectal cancer

- λ_S estimated from familial aggregation as 1.7:
  - optimal info for discrimination 0.77 bits
- 3689 cases, 12349 controls in UK Biobank
- Locus-specific polygenic scores calculated on cases and controls using summary results from meta-analysis of GWAS for colorectal cancer
- LASSO regression of colorectal cancer on scores
  - predictive performance evaluated by cross-validation
- Polygenic score contributes only 0.1 bits of information for discrimination

Incremental contribution of microbiome profile to detection of colorectal cancer

Model	Cases / controls	C-statistic	Average weight of evidence (bits)	Test log-likelihood (bits)
FIT_only	101 / 141	0.894	2.99	132.5
FIT+microbiome	101 / 141	0.928	6.55	86.2

Genetic scores: prediction of drug response in rheumatoid arthritis

- From SNP relationship matrix, genetic factors account for up to 30% of variance in respone response to anti-TNF therapy
- GWAS studies do not find any hits
- GENOSCORES platform: database of summary GWAS results on clinical traits and biomarkers
    - can compute genotypic scores for each locus and each trait in a target genotype dataset
    - these scores can be used as genotypic features to predict outcome
    - associations with scores may be detectable when associations with SNPs are not: smaller prior hypothesis space.
- MATURA collaboration: Response to anti-TNF agents measured in 3294 people with rheumatoid arthritis
- Scores computed for rheumatoid arthritis, immune cell traits cell frequency and surface protein expression), expression of genes previously associated with drug response.

Can associations with genotypic scores be detected where there are no GWAS hits?

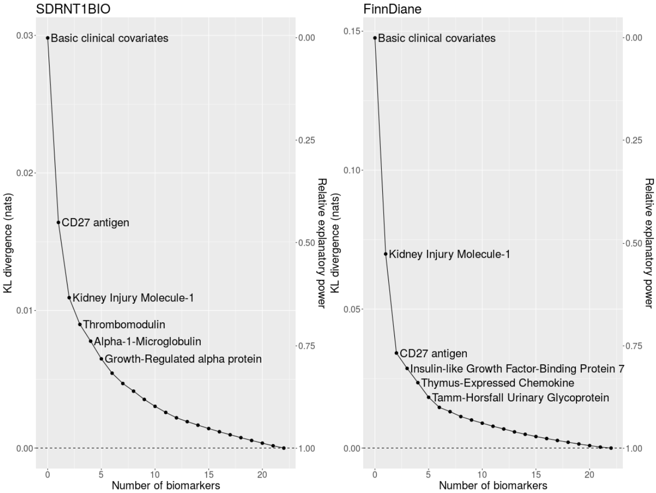
Trait scores	Δ log L	variance explained
eQTLs	4.1	0.2%
Rheumatoid arthritis	2.5	0.1%
Immune cell	2.1	0.1%

- strongest RA score effect at *CD40* locus
- stronges immune cell trait at ENTPD1 locus coding for CD39 protein which mediates anti-inflammatory effects of methotrexate
    - low expression of CD39 on T regulatory cells previously associated with poor response to RA

Projective predictive selection

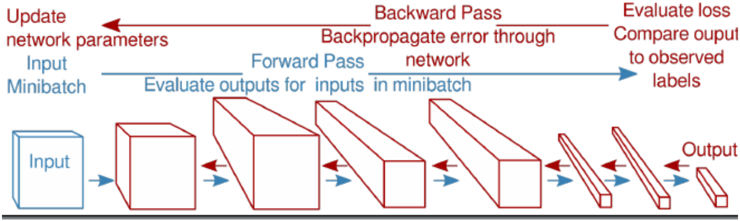
- Stan allows us to evaluate the predictive performance of an entire biomarker panel, learning the distribution of effect sizes from the data
- Having evaluated how well a biomarker panel can predict outcome, we usually want to choose the smallest subset of biomarkers that will contain most of the predictive information
- *Projective predictive selection* allows us to do this without re-using the test data.
    - Given posterior samples of the linear predictor, compute regression of predicted value on biomarkers added one at a time by forward selection

Projective predictive selection of biomarkers for progression of diabetic nephropathy



Classifying images using using deep learning

- Deep learning outperforms all other algorithms for computer vision, speech recognition
  - not really artificial intelligence
- Convolutional neural network: overlapping patches of pixels in image are passed through layers of generalized linear models

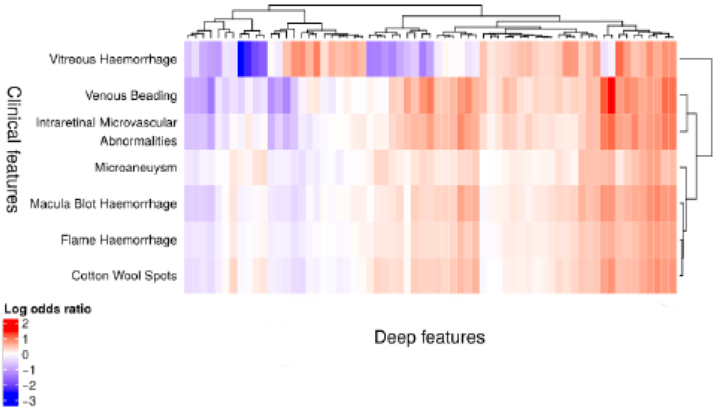


Time to referable retinopathy: Increment in predictive performance when adding deep learning to baseline model (clinical data + manual grading)

- 30,604 manually graded retinal images from 3290 people in Scottish Type 1 Diabetes Bioresource
    - linked to clinical data for years 2007 to 2016.
- 384 features derived from CNN trained on a public dataset calculated on these images.
- Generalized linear model with clog-log link function
    - forward selection compared with Bayesian model with horsehoe prior

Model	C-statistic	Λ (bits)
Baseline	0.81	0.75
Baseline + CNN (forward-selected)	0.88	2.5
Baseline + CNN (horseshoe-prior)	0.91	3.0

Relation of features extracted by deep learning to clinical observations



Regulation and impact on health services

- FDA 2007 draft guidelines, withdrawn 2010
  - univariate “laboratory-developed tests” are subject to “enforcement discretion”
  - FDA enforces lab quality on these, but leaves interpretation to the clinician
  - In vitro diagnostic “multivariate index” assays are not transparent to the clinician and should be regulated by FDA.
  - prospective studies preferred, but retrospective studies using archived samples may sometimes be used
- FDA current regulations:
  - Level 1 evidence required for a **companion diagnostic** that provides information essential for safe and effective use of a therapeutic product.
  - Less stringent Level 2 evidence required where health professionals can use their own judgement

Predictive and prognostic biomarkers?

- FDA 2016:
  - predictive biomarker: treatment × biomarker interaction effect on outcome
  - prognostic biomarker: average effect of biomarker on outcome
- prognostic biomarkers influence decision to treat, but not to which treatment to use
- For most diseases, large genotyped cohorts including treatment allocation and standardized measurements of outcome are not available.

Do we need randomized trials to evaluate predictive biomarkers?

- Drug effects in observational studies are heavily confounded by unmeasured factors that influence treatment allocation and with outcome
  - observational studies are likely to give wrong answers (for instance with effect of post-menopausal oestrogens on cardiovascular disease)
- Treatment x biomarker interaction effects are not in general confounded, unless the biomarker is associated with an observable clinical trait.
- Evidence for prognostic biomarkers does not in general require randomization
- Precision medicine-based diagnosis should be a continuous learning process, not tests that are cast in stone once evaluated.

Can precision medicine reduce drug cost?

- Value-based pricing in the UK
  - National Institute for Clinical Excellence negotiates drug price based on cost per quality-adjusted life-year gained
  - Identifying the subset of patients in whom the drug is effective will reduce the number of patients treated with a given drug but the value added per treated patient will rise
- More selective use of drugs will not reduce their costs: R & D costs of bringing a drug to market (about 1% of GDP in OECD countries) have to be covered.
- Pricing per dose is economically inefficient.
  - marginal cost of extra dose is low, but each dose is priced to recover development costs
- Possible alternatives:
  - national licensing at national level for unlimited use
  - governments buy patent rights to molecules
  - private sector competes for government funding of drug development

Conclusions

1. Learning to subtype disease is possible in principle but needs better tools for statistical computation
2. To quantify performance of a classifier, $C - statistic$ should be replaced by average weight of evidence.
3. Prediction of drug response from biomarkers at baseline is not looking promising
    - surrogate measures of drug response are more likely to be useful
    - Genetic biomarker studies may be good science, but not for use as clinical predictors anyway
4. Precision medicine does not need randomized evaluation, and will not reduce drug costs