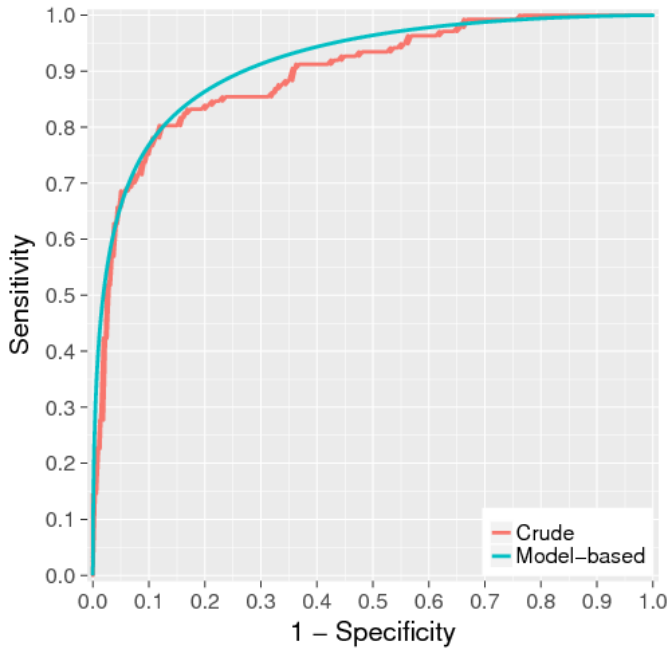
Quantifying predictive performance using the distributions of weight of evidence

Paul McKeigue

Usher Institute of Population Health Sciences and Informatics

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

Problems with the C-statistic

- No obvious application to risk stratification
- Not useful for quantifying incremental contribution of a new predictor to a baseline model
  - 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.
- Not easily extended to survival data
- Crude ROC curve (based on comparing all possible case-control pairs) is not necessarily concave.
  - For any ROC curve that is not concave, one can produce a classifier superior to that summarized by such an ROC curve (Hand, 2009).

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

"Identifying suitable measures for quantifying the incremental value of adding a predictor to an existing prediction model remains an active research area" (Collins 2015).

Bayesian approach to hypothesis testing and classification

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

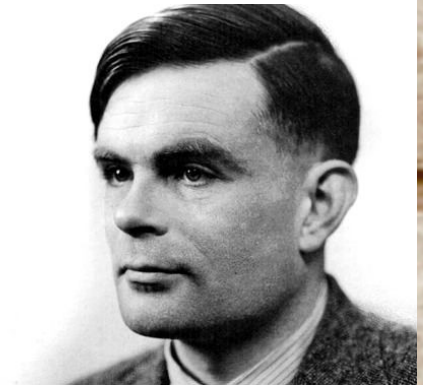
$$(\text{prior odds } \mathcal{H}_1 : \mathcal{H}_0) \times \frac{\text{likelihood of } \mathcal{H}_1}{\text{likelihood of } \mathcal{H}_0} = (\text{posterior odds } \mathcal{H}_1 : \mathcal{H}_0)$$

The ratio of likelihoods of hypotheses is the **Bayes factor**.
Taking logarithms, this becomes

log prior odds $\mathcal{H}_1/\mathcal{H}_0$ + **weight of evidence** $\mathcal{H}_1/\mathcal{H}_0$ = log posterior odds $\mathcal{H}_1/\mathcal{H}_0$

- 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

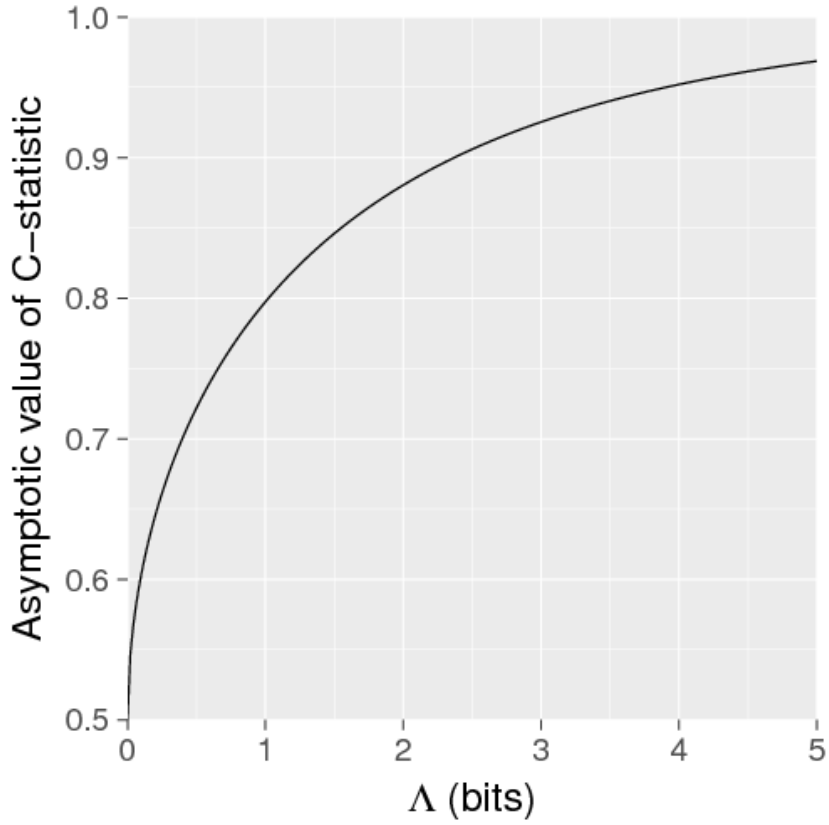




Banburismus procedure accumulated weights of evidence for settings of right and middle rotors of the Enigma machine
Turing: if the sampling distribution of the weight of evidence W in favour of a hypothesis when it is true is gaussian with mean Λ:

- distribution of W when the hypothesis is false is gaussian with mean $-\Lambda$
- both distributions have variance 2Λ (when natural logarithms are used)

Distribution of W will be gaussian if there are many independent predictors of small effect.

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 to 0.8, or from 0.88 to 0.925

General relationship between distributions of weight of evidence in cases and controls (Good and Toulmin, 1968)

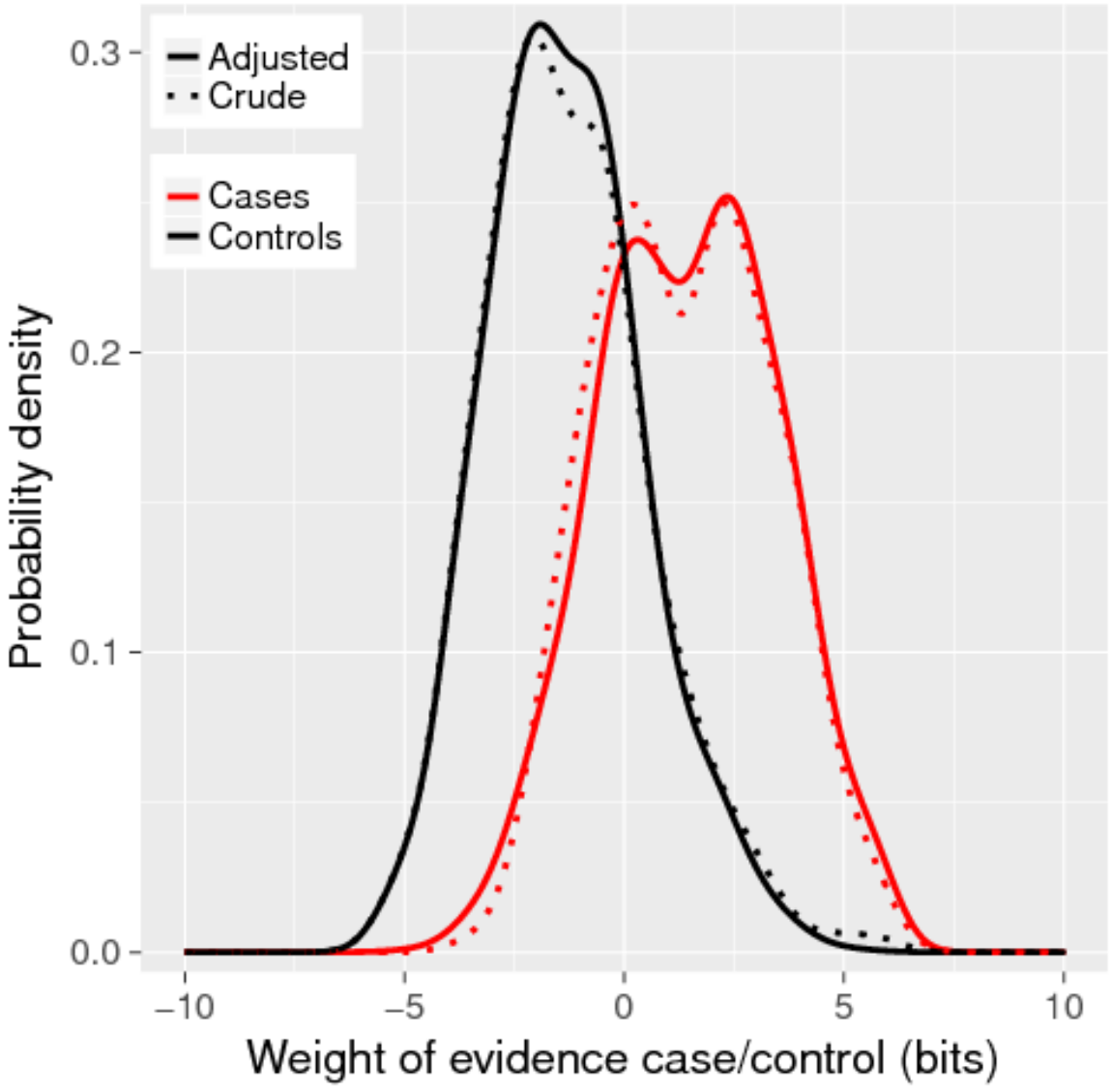
Write

- W for weight of evidence favouring $\mathcal{H}_1$ over $\mathcal{H}_0$
- $p_1(W)$ for density of W when $\mathcal{H}_1$ is true
- $p_0(W)$ for density of W when $\mathcal{H}_0$ is true

At any value of W the ratio $p_1(W)/p_0(W)$ is the Bayes factor $\exp(W)$ favouring $\mathcal{H}_1$ over $\mathcal{H}_0$.

$$\exp(-W)p_1(W) = p_0(W)$$

Distributions in cases and controls of weight of evidence W favouring case over control status



Calculating weight of evidence W favouring case over control status on test data

Use model to compute on each test observation:

- posterior probability of case status
- prior probability of case status (from null model)
- Calculate $W = \log$ posterior odds - log prior odds

Can extend this to survival analysis

Other names for expected weight of evidence Λ

- *Expected information for discrimination* between cases and controls
- Kullback-Leibler (KL) divergence from the class-conditional distribution $\mathcal{Q}$ of the predictors under incorrect case-control assignment to their distribution $\mathcal{P}$ under correct assignment
- relative entropy of $\mathcal{P}$ with respect to $\mathcal{Q}$.

As Λ is a KL divergence, it can take only non-negative values.

Advantages of using expected weight of evidence Λ to quantify performance of a diagnostic test

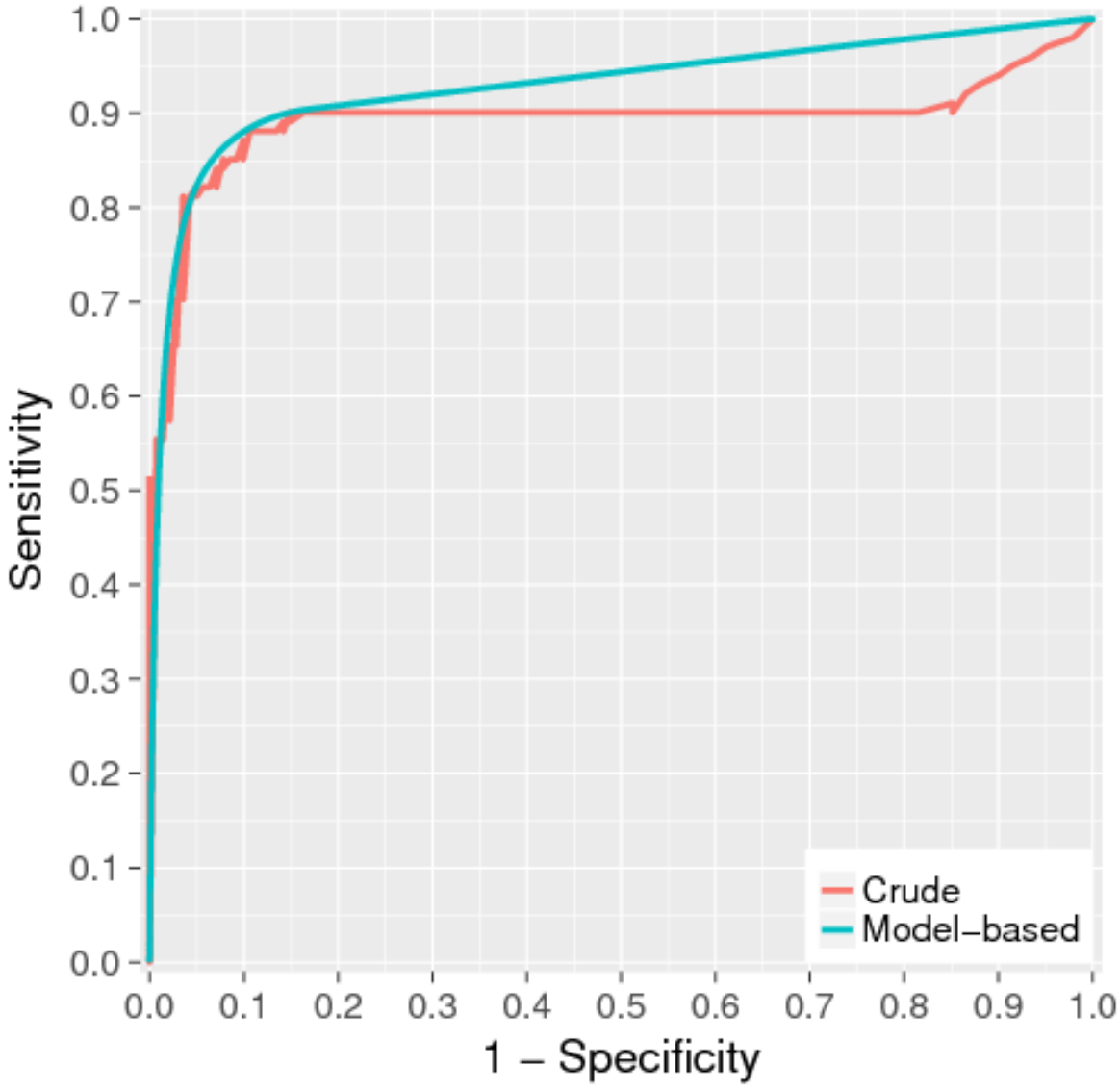
- Contributions of independent variables to predictive performance are additive on the scale of Λ.
- Expected weight of evidence has an intuitive interpretation as the typical factor by which prior odds are updated to posterior odds
- Where there are many independent predictors of small effect, expectation of the weight of evidence determines its asymptotic distribution and this contains all the information required to characterize fully how the test will behave as a risk stratifier.
- Calculation of weight of evidence can be extended to interval-censored failure-time data

Relation of ROC curve to distributions of weight of evidence

Johnson (2004): model-based ROC curve can be calculated from the distributions of W in cases and controls:

- if the quantiles of W in controls and cases are q_0 and q_1 respectively, the sensitivity is $(1 - q_1)$ and the specificity is q_0, the ROC is the curve obtained by plotting $(1 - q_1)$ as a function of $(1 - q_0)$.
- The gradient of this model-based ROC curve is the Bayes factor $\exp(W)$
- Model-based ROC curve is concave (downwards)

Comparison of crude and model-based ROC curves for prediction of colorectal cancer from FIT test



What value of Λ corresponds to useful prediction?

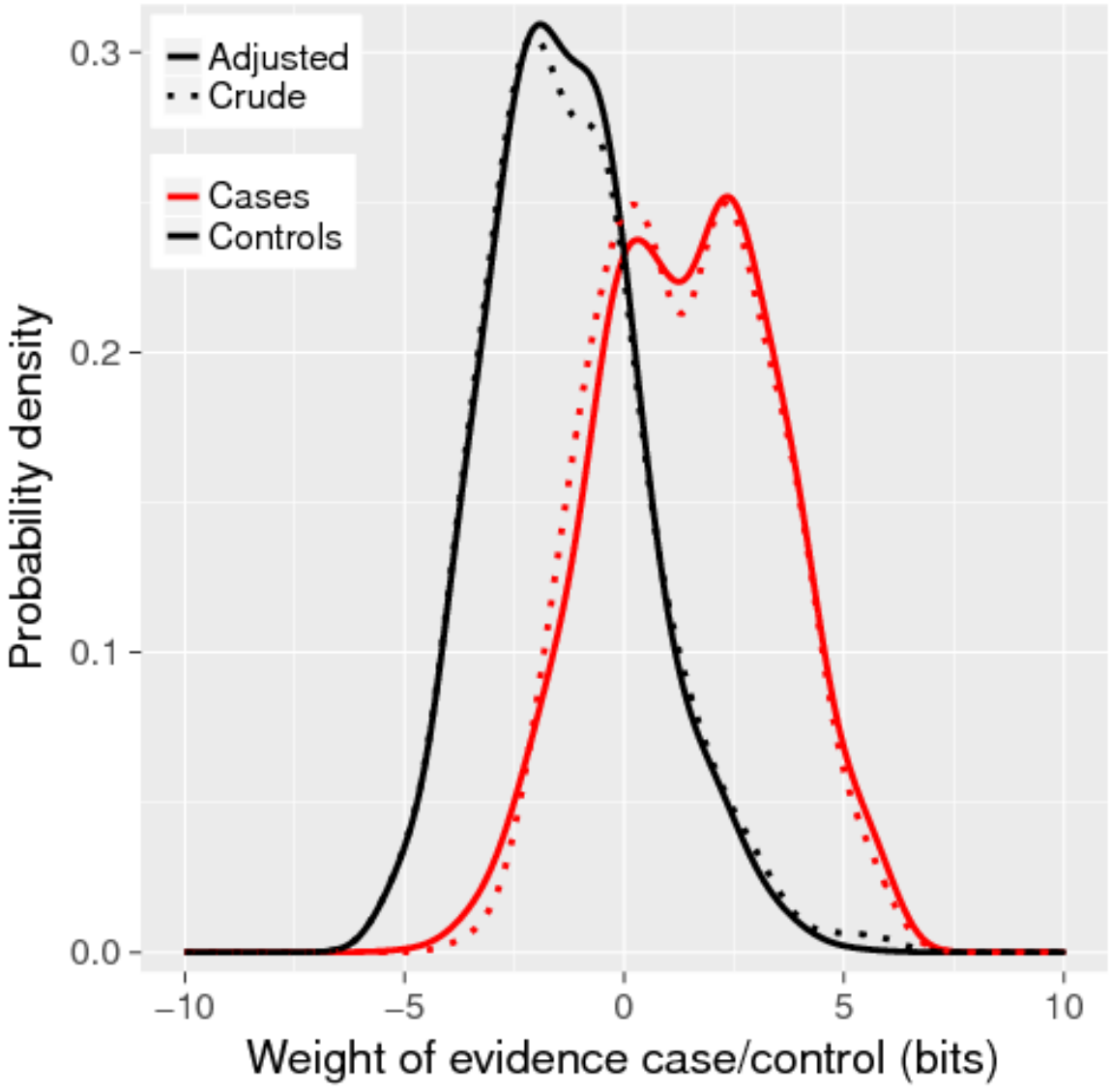
- Suggested criteria in a clinical setting:-
  - Moderate performance: 1 bit (C=0.80)
  - Good performance: 3 bits (C=0.925)
- For population screening:
  - Moderate performance: 3 bits (e.g. FIT testing for colorectal cancer)
  - Good performance $>$ 5 bits
- Even a good test will often give wrong answers: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

Diabetes in Pima Americans

Table 1: Prediction of diabetes in Pima Native American women

Model	Cases / controls	C-statistic	Average W in cases (bits)	Average W in controls (bits)
Pima_diabetes	268 / 500	0.838	1.39	-1.28

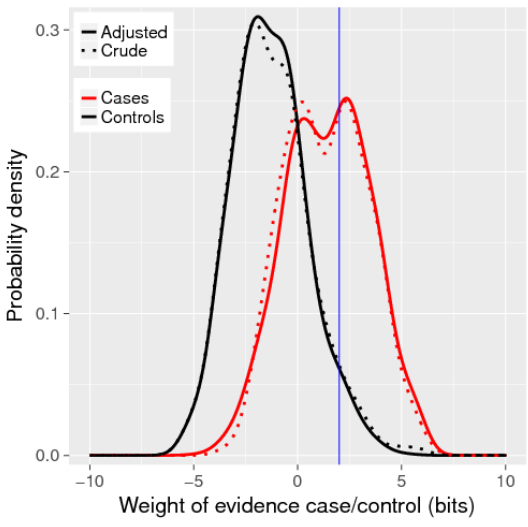
Distributions in cases and controls of weight of evidence W favouring case over control status

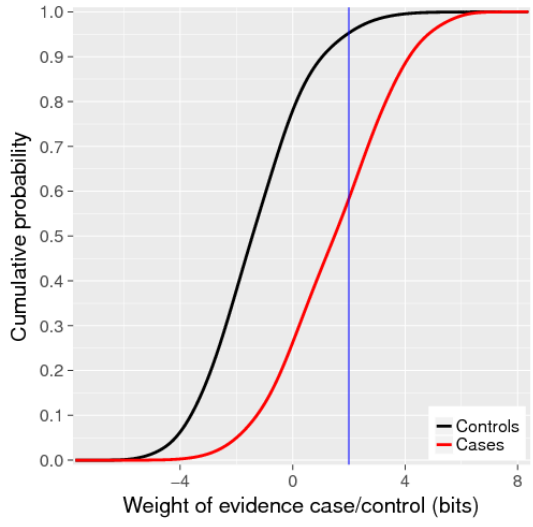


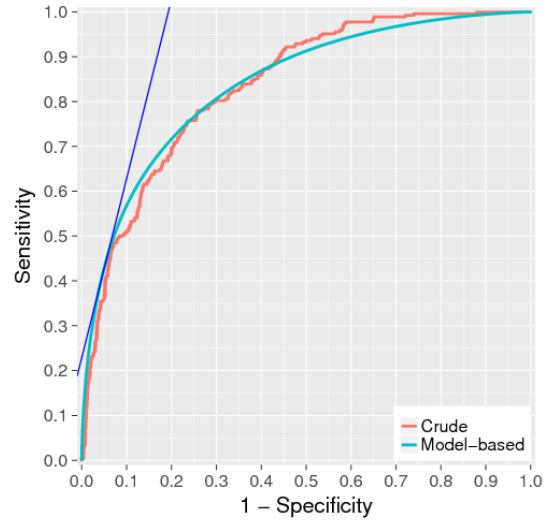
Risk stratification: relation of model-based ROC curve to distribution of W

Example: cutoff for risk stratification is a Bayes factor of 4 (blue line at $W = 2$)

- 95% of controls and 58% of cases (95% specificity, 42% sensitivity) are below this threshold.
- Gradient of model-based ROC curve is Bayes factor







Conclusions

- To report performance of a diagnostic classifier, plot the distributions of W in cases and controls and summarize predictive performance as Λ:
  - average weight of evidence favouring true over false status
  - evaluated as average of the means in cases and controls, in bits
- If you have to show an ROC curve, show the crude and model-based curves
- For survival data, use a null model (prediction from interval length only) to calculate prior probs on test dataset
- R package `wevid` provides functions for all required calculations and plots