# Evaluating performance of a diagnostic test as the expected weight of evidence

**Paul McKeigue**

Usher Institute of Population Health Sciences and Informatics, University of Edinburgh

## Introduction

The paper **Quantifying performance of a diagnostic test as the expected information for discrimination: relation to the C-statistic**, published in *Statistical Methods for Medical Research* (2018), has attracted wide interest. This poster explains the ideas.

THE TIMES

Turing's work could help with enigma of cancer detection

Katrine Bussey

Work by the Second World War code-breaker Alan Turing could help to develop better tests for the early detection of cancer and other diseases.

Researchers at the University of Edinburgh believe his mathematical techniques could be used to help to measure the effectiveness of diagnostic tools.

The accuracy of diagnostic tests is assessed using statistical techniques developed in the 1980s, but these are unable to gauge how useful a test could be in determining an individual's risk of developing a disease.

Scientists at the university's Usher Institute of Population Health Sciences and Informatics believe that Turing's methods could improve the techniques.

Turing's mathematical genius played a key role in breaking the Nazi Enigma code. Working at Bletchley Park —

The Daily Telegraph

Turing method that cracked Enigma to help in cancer fight

*By* **Henry Bodkin**

THE method devised by Alan Turing, the Second World War codebreaker, to crack Enigma could be used to detect cancer earlier, experts have said.

Researchers at Edinburgh University believe Turing's mathematical techniques could be used to help measure the effectiveness of existing diag-

www.news.cn 新华网 NEWS www.xinhuanet.com

Legendary codebreaker's legacy can still help modern medical development

LONDON, Oct. 30 (Xinhua) -- By building on the unpublished work of Second World War codebreaker Alan Turing, researchers may find a new way to enhance early diagnosis of cancer and emerging diseases, according to a new study recently released by the University of Edinburgh.

Working at Bletchley Park in 1941, Turing devised his method used to break the German forces' Enigma code. Turing's approach investigated the distribution of so-called weights of evidence - which establish the likely outcomes in a given situation - to help him decide the best strategy for breaking Enigma.

Research at the University of Edinburgh suggests an approach, based on Turing's method, could test the accuracy of diagnostic tools.

Professor Paul McKeigue, of the University's Usher Institute of Population Health Sciences and Informatics, shows that the same principle of how the weight of evidence varies can be applied to evaluate the diagnostic tests

## Evaluating diagnostic tests

### Why we need better methods

"Precision medicine" presents new challenges. Clinical leads, regulatory agencies and industry need to evaluate:

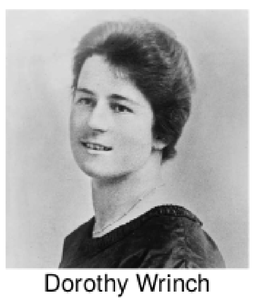
- how much extra information is gained by adding a new test to existing diagnostics.
- how a test will perform for risk stratification: proportions of true and false positives at a given risk threshold.

### Limitations of the $C$-statistic

Current methods are based on the $C$-statistic: area under the Receiver Operating Characteristic (ROC) curve:

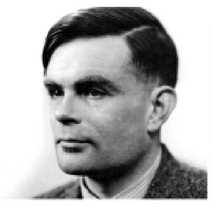
- increment in $C$-statistic obtained by adding a new predictor is difficult to interpret
- cannot calculate proportions of true and false positives at a given risk threshold.

## Historical note



Dorothy Wrinch

Hut 8, Bletchley Park



Alan Turing

Jack Good

**Dorothy Wrinch** and **Harold Jeffreys** (1921) were the first to write Bayes theorem in the odds form, showing that the ratio between likelihoods of hypotheses, later called the **Bayes factor**, transforms prior odds into posterior odds. Taking logarithms, we can write this equation in terms of the **weight of evidence** (log Bayes factor).

$$\log \text{prior odds } \mathcal{H}_1 : \mathcal{H}_2 + \text{weight of evidence } \mathcal{H}_1 : \mathcal{H}_2 = \log \text{posterior odds } \mathcal{H}_1 : \mathcal{H}_2$$

The *Banburismus* procedure devised by **Alan Turing** at Bletchley Park was based on accumulating weights of evidence for the settings of the Enigma machine. **Jack Good**, who was Turing's assistant at Bletchley Park, recounted in 1994: 'One morning I asked Turing "Isn't this really Bayes' theorem?" and he said "I suppose so."'

## Statistical properties of the weight of evidence

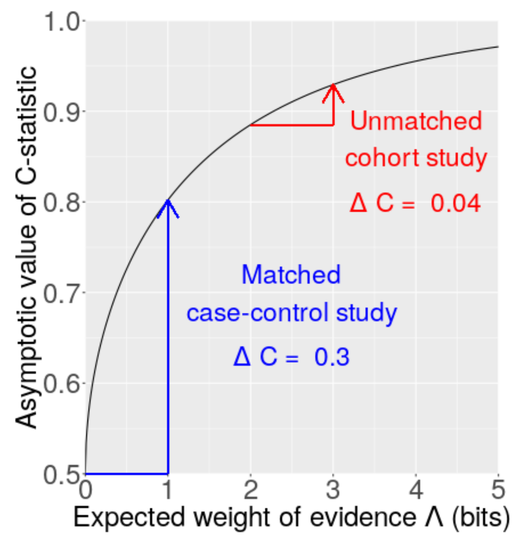
To predict whether Banburismus would work, Turing in 1940 began investigating the sampling distribution of the weight of evidence $W$. He discovered two key results:
1. If the density $p(W)$ of the weight of evidence $W$ favouring a given hypothesis when it is true is Gaussian with mean $\Lambda$, the density $q(W)$ when that hypothesis is false is Gaussian with mean $-\Lambda$, and both densities have variance $2\Lambda$.
2. The expected Bayes factor in favour of a hypothesis when it is false is 1. A corollary is that when a good test is wrong, it will often be wildly wrong.
Good and Toulmin (1968) extended these results by showing that even when the densities $p(W)$ and $q(W)$ are not Gaussian, there is a mapping between their respective characteristic functions $\phi_{\text{true}}(t)$ and $\phi_{\text{false}}(t)$ given by the identity $\phi_{\text{false}}(t) = \phi_{\text{true}}(t + i)$.

## Why use expected weight of evidence $\Lambda$?

- $\Lambda$ is the **expected information for discrimination** between cases and controls. The mathematical definition of information corresponds to intuitive ideas of information:
  - surprising observations convey more information than commonplace ones
  - information contributed by independent tests can be added.
- Increment in $\Lambda$ does not depend on covariates that are independent of the new predictor
- $\Lambda$ can be extended to screening programmes (interval-censored data).
  If we use logarithms to base 2, the weight of evidence can be expressed in *bits*, which have a more intuitive interpretation than natural log units.

## Relation of $\Lambda$ to C-statistic

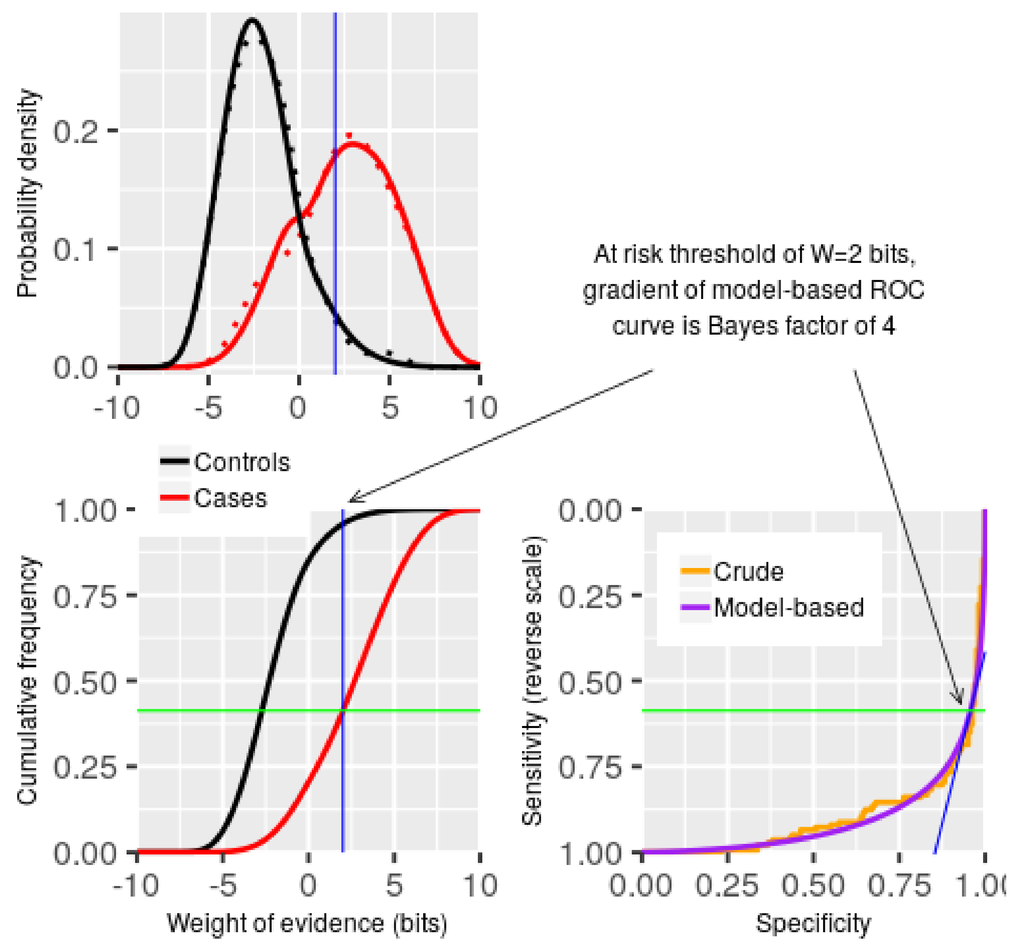
If there are many independent predictors of small effect, the weight of evidence will have the distribution derived by Turing and the $C$-statistic can be interpreted as a mapping of the expected weight of evidence $\Lambda$, which takes values from 0 to infinity, to the interval from 0.5 to 1.
From this curve we can see why increments in $C$-statistic are hard to interpret: for a biomarker contributing expected weight of evidence of 1 bit, the increment in $C$-statistic will be larger in a case-control study in which covariates have been matched than in a cohort study in which these covariates contribute a weight of evidence of 2 bits.




Asymptotic relationship of $C$-statistic to expected weight of evidence $\Lambda$

## Example: risk stratification

The R package *wevid* (available on CRAN) estimates the densities $p(W)$ and $q(W)$, of the weight of evidence $W$ in cases and controls, subject to the constraint of the mapping between these densities. The plots below show these estimated densities in a study of predictors of coronary disease in Cleveland, together with cumulative frequency distributions, and the model-based ROC curve (with axes reversed) calculated from these frequency distributions. The expected weight of evidence $\Lambda$ is 2.3 bits.




Suppose we set a risk threshold based on a Bayes factor of 4 ($W = 2$ bits, vertical blue line). At this threshold, 96% of controls and 41% of cases will be excluded. At this sensitivity (horizontal green line), the gradient (blue tangent line) of the model-based ROC curve is the Bayes factor. The model-based ROC curve encodes the same information as the cumulative frequency distributions but is harder to use for risk stratification because the Bayes factor cannot be read from the axis.