Determining cut-points for Alzheimer's disease biomarkers: statistical issues, methods and challenges

Jonathan W. Bartlett[1,2], Chris Frost[1,2], Niklas Mattsson[3], Tobias Skillbäck[3], Kaj Blennow[3], Henrik Zetterberg[2,3], Jonathan M Schott[2*]

1 Centre for Statistical Methodology, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

2 UCL Institute of Neurology, Queen Square, London, WC1N 3BG, UK

3 Institute of Neuroscience and Physiology, Department of Psychiatry and Neurochemistry, The Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden

* To whom correspondence should be addressed

Abstract

New proposed criteria for the clinical diagnosis of Alzheimer's disease (AD) increasingly incorporate biomarkers, most of which are normally measured on a continuous scale. Operationalizing such criteria thus requires continuous biomarkers to be dichotomised, which in turns requires selection of a cut-point at which to dichotomise. In this article we review the statistical principles underlying the choice of cut-point; describe some of the most commonly adopted statistical approaches used to estimate cut-points; highlight potential pitfalls in some of the approaches; and characterise in what sense the estimated cut-point from each approach is optimal. We also emphasize that how a cut-point is selected must be made in reference to how the resulting dichotomised biomarker is to be used, and in particular what actions will follow from a positive or negative test result.

Keywords: biomarker, dichotomisation, accuracy, sensitivity, specificity, cerebrospinal fluid.

Introduction

With the prospect of disease modifying therapies for Alzheimer's disease (AD) there is an urgent requirement for accurate and ever earlier diagnosis. Improvements in our understanding of the biology of AD have led to remarkable advances in clinical practice – perhaps exemplified by the now widespread use of cerebrospinal fluid (CSF) biomarkers including Aβ1-42 which is decreased in AD; and total tau (t-tau) and phosphorylated tau (p-tau), which are both typically increased in AD. Studies have demonstrated that these biomarkers are associated with clinical (1) and pathologically proven AD (2); can help determine which subjects with mild cognitive impairment will develop dementia due to AD (3,4); and perhaps may even be able to identify apparently healthy elderly individuals in the earliest stages of neurodegeneration (5). The now widespread view that the pathological changes of AD start many years prior to the onset of dementia (6) and that this pre-dementia phase would be the ideal time to instigate disease-modifying therapies, has still further intensified the search for disease specific biomarkers to identify at-risk individuals.

Biomarkers – including magnetic resonance imaging (MRI), amyloid positron emission tomography (PET) and cerebrospinal fluid (CSF) measures – are now being incorporated into proposed diagnostic criteria (7,8). Thus for example the most recent proposed criteria for preclinical AD – itself divided into three stages – all require evidence for "amyloid positivity" based either on CSF or amyloid PET imaging (8). Similar requirements are increasingly needed in the context of clinical trials, where evidence of "amyloid positivity" may be required as an entry criterion; for a range of research studies where assigning individuals as positive/negative on a range of biomarkers may be helpful in understanding the pathogenesis and progression of the disease; and ultimately even perhaps for screening at-risk healthy populations.

As acknowledged in the paper in which the preclinical AD criteria are proposed (8), defining amyloid positivity, or for that matter biomarker positivity in general is not straightforward, and explicitly requires the determination of a biomarker "cut-point" to distinguish amyloid-positive from amyloid-negative individuals. Dichotomising a continuous biomarker as normal/abnormal according to a pre-defined cut-point simplifies its interpretation, at the inevitable expense of loss of information. How a cut-point is chosen depends critically on how and in which population the resulting dichotomised biomarker is to be used. Furthermore, there are a wide variety of different statistical approaches which can potentially be used to find a cut-point. For both these reasons, for any given biomarker, a single optimal cut-point does not exist.

In this article, we review some of the issues involved in determining cut-points, and describe some of the most commonly used statistical approaches for estimating "optimal" cut-points for an AD biomarker. We start by describing some of the different contexts in which dichotomised biomarkers might be used, as different scenarios demand different approaches to cut-point selection. We then review the key statistical concepts which characterise the performance of a binary test (e.g. a dichotomised continuous biomarker), before describing some of the most commonly adopted statistical methods for determining cut-points. Whilst for the most part we discuss issues in general terms, we illustrate some of the methods and issues using data from Mattsson *et al* (3). The methods by which these data were acquired and analysed have previously been described in detail elsewhere (3), but in brief measures of Aβ1-42, t-tau and p-tau from 529 patients with a clinical diagnosis of AD and 324 controls from a multi-centre CSF biomarker study were available for analysis.

Applications of dichotomised AD biomarkers

Before describing the various statistical approaches for choosing cut-points, we must first consider the potential uses of dichotomised AD biomarkers. Perhaps the most common is when the

dichotomised biomarker is to be used to make or aid clinical diagnoses. Dichotomised biomarkers might also be used either in isolation or in combination with other biomarkers for screening populations, i.e. to identify individuals who are at high risk of future development of AD. Inclusion/exclusion criteria for randomized clinical trials are typically based on a number of dichotomized biomarkers. Biomarkers are also often dichotomised for the purposes of simplifying statistical analyses in research studies. As we describe in the remainder of this paper, choosing a cut-point in order to dichotomise a biomarker involves a trade-point between falsely classifying diseased (or future diseased) subjects as non-diseased and vice-versa. How this trade point is made depends critically on the intended use of the dichotomised biomarker, the population in which it is to be used, and the relative costs of making the two types of errors. It is thus essential to bear in mind that what may be an appropriate cut-point for a particular biomarker in one application and population may be inappropriate in a different setting.

Statistical properties of a dichotomised test

Terminology

To define the key statistical quantities that describe the performance of a dichotomised biomarker we use X to denote a measurement from the biomarker in question, and assume that higher values of X are associated with disease. We use the letter D to indicate whether a subject has the disease (D=1) or not (D=0). In this terminology, in some situations (e.g. screening apparently healthy populations) D may refer to whether or not a subject will subsequently develop the disease of interest at some time in the future, as opposed to their disease status when X is measured. Our goal is to choose a cut-point, denoted c, which will be used to dichotomise the biomarker. Thus subjects with $X \geq c$ will be classified as "positive", and those with $X < c$ as "negative".

Sensitivity and specificity

For a given cut-point value c, the sensitivity of the "test" (i.e. using the dichotomised biomarker to classify subjects or to make a decision of some kind) is the proportion of truly diseased subjects (cases) who have a positive test result, i.e. in probability notation $P(X \geq c|D=1)$ (the probability that a randomly chosen diseased subject has $X \geq c$). The specificity of the test is the proportion of non-diseased subjects who have a negative test result, denoted $P(X<c|D=0)$ (Table 1). Since sensitivity and specificity are quantities that condition on disease status, they can be estimated from case-control studies (studies which sample according to disease status).

Table 1.

Numbers (or proportions) of subjects classified according to disease status (D) and dichotomised (at cut-point c) biomarker (X).

	Biomarker positive ($X \geq c$)	Biomarker negative ($X<c$)
Subjects with the disease (D=1)	α	γ (False negative)
Subjects without the disease (D=0)	β (False positive)	δ

$$\left.\begin{aligned} \text{Sensitivity} &= P(X \geq c|D=1) = \frac{\alpha}{\alpha + \gamma} \\ \text{Specificity} &= P(X<c|D=0) = \frac{\delta}{\beta + \delta} \end{aligned}\right\}$$ Appropriate even if the proportion of subjects with disease is not the same in the sample and the target populations

Positive Predictive Value (PPV)	$= P(D=1\|X\geq c) = \frac{\alpha}{\alpha + \beta}$	Only appropriate if the proportion of subjects with disease is the same in the sample and the target populations
Negative Predictive Value (NPV)	$= P(D=0\|X<c) = \frac{\delta}{\delta + \gamma}$	

Positive and negative predictive value

The positive predictive value (PPV) is the probability that a subject has the disease given that his or her test result is positive. Algebraically, this is $P(D=1|X\geq c)$. If data are sampled in such a way that the proportion of disease subjects is representative of the target population, this can be estimated by the proportion of biomarker positive subjects who have the disease (Table 1). The negative predictive value (NPV) is the probability that a subject given a negative test result is truly not diseased: $P(D=0|X<c)$, or $\delta/(\delta + \gamma)$ in Table 1.

Given a subject's test result, positive and negative predictive values inform the likelihood that the subject has the disease. It is thus of direct interest when a subject is tested but disease status is unknown. This is in contrast to sensitivity and specificity, which give probabilities conditional on disease status. In most applications where dichotomised biomarkers might be used, the test result is known and disease status is unknown (hence the use of the biomarker). Therefore predictive values are usually of more interest to the clinician/patient than sensitivity and specificity. Probability theory shows that positive and negative predictive values are equal to (9):

$$\mathrm{PPV} = \frac{\mathrm{Prevalence} \times \mathrm{Sensitivity(c)}}{\mathrm{Prevalence} \times \mathrm{Sensitivity(c)} + (1\text{-}\mathrm{Prevalence}) \times (1\text{-}\mathrm{Specificity(c)})} \quad [1]$$

$$\mathrm{NPV} = \frac{(1 - \mathrm{Prevalence}) \times \mathrm{Specificity(c)}}{(1 - \mathrm{Prevalence}) \times \mathrm{Specificity(c)} + \mathrm{Prevalence} \times \big(1 - \mathrm{Sensitivity(c)}\big)} \quad [2]$$

Predictive values can thus only be estimated using data in which the proportion of diseased subjects is representative of the target population or if an external estimate of prevalence is available. That predictive values depend on prevalence explains the fact that a test can have very high sensitivity and specificity, yet low positive predictive value in a population with low prevalence (10).

Accuracy

In the context of binary tests, the term accuracy usually refers to the probability of giving the correct test result (i.e. a positive result to cases and a negative test result to controls). Probability theory shows accuracy can be expressed as:

$$\mathrm{Accuracy(c)} = \big(\mathrm{Prevalence} \times \mathrm{Sensitivity(c)}\big) + \big((1 - \mathrm{Prevalence}) \times \mathrm{Specificity(c)}\big) \quad [3]$$

Thus, as for predictive values, accuracy is also dependent on disease prevalence. Thus a test might be deemed "accurate" in one population but "inaccurate" in another population with differing disease prevalence.

Likelihood ratios

In contrast to the above measures, positive and negative likelihood ratios are not dependent on disease prevalence. The positive likelihood ratio (LR+) is the ratio of sensitivity to 1-specificity whilst the negative likelihood ratio (LR-) is the ratio of 1-sensitivity to specificity. Their importance is seen when equations [1] and [2] are rearranged to express relationships between odds of disease rather than risk (odds = risk/(1-risk)). The formulae become:

$$\frac{\text{PPV}}{1-\text{PPV}} = \frac{\text{Sensitivity(c)}}{1-\text{Specificity(c)}} \times \frac{\text{Prevalence}}{1-\text{Prevalence}} = (\text{LR}+) \times \frac{\text{Prevalence}}{1-\text{Prevalence}} \quad [4]$$

$$\frac{1-\text{NPV}}{\text{NPV}} = \frac{1-\text{Sensitivity(c)}}{\text{Specificity(c)}} \times \frac{1-\text{Prevalence}}{\text{Prevalence}} = (\text{LR}-) \times \frac{1-\text{Prevalence}}{\text{Prevalence}} \quad [5]$$

Equation [4] shows that LR+ is a multiplicative factor that acts on the odds of disease in those who are test positive. For example if the sensitivity is 75% and the specificity is 97% the LR+ is 25 (75/3) and so individuals who have positive test results have odds of disease that are 25 times that in the population taking the test. The LR+ and LR- typically both monotonically increase with increasing cut-off meaning that in themselves they are not useful for selecting cut-points, the topic we turn to next.

Statistical methods for choosing cut-points

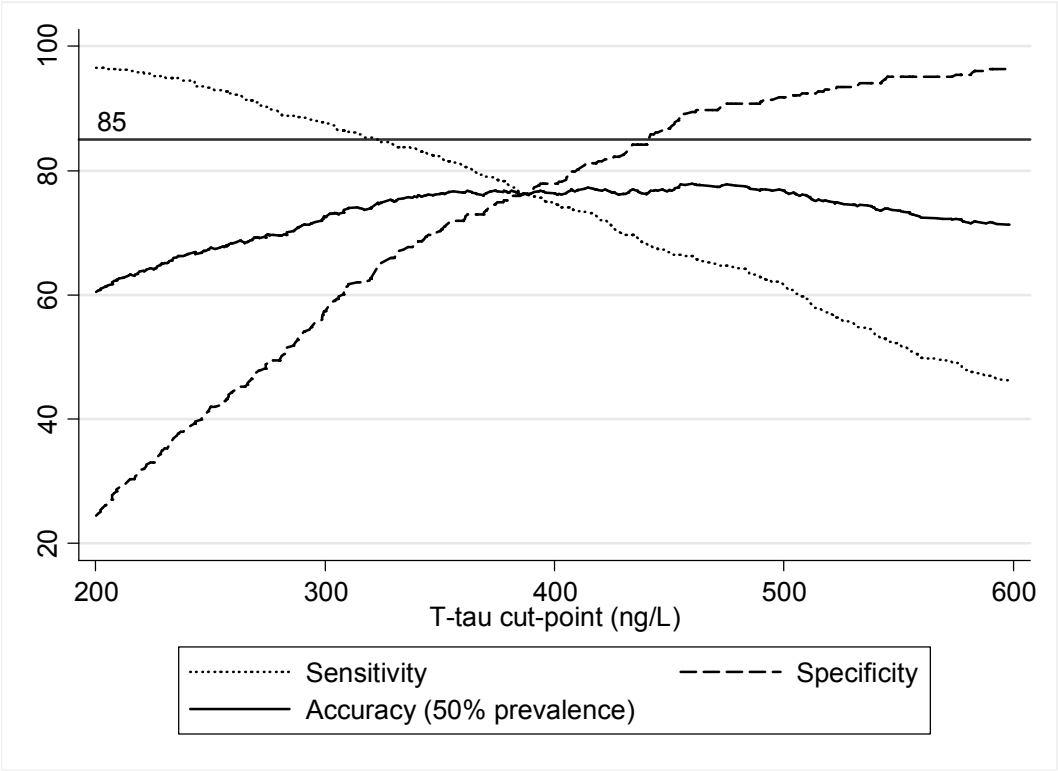
We now turn to the problem of estimating or choosing the cut-point c to turn a continuous variable into a dichotomised test. A perfect test has 100% sensitivity and 100% specificity for one or more cut-points. Alas tests are rarely, if ever, perfect. In practice therefore, how we choose the cut-point c affects the sensitivity and specificity of our test. As depicted in Figure 1, which shows the sensitivity and specificity of t-tau for diagnosing clinical AD estimated using data from Mattsson *et al* (3), decreasing c means that more diseased subjects will (correctly) be given positive test results, thus increasing sensitivity. However, this leads to a reduction in specificity. Conversely, increasing the cut-point c will increase specificity, but reduce sensitivity. In choosing a cut-point c there is thus an inevitable trade-point to be made between which type of error (wrongly classifying a control as a case, and vice-versa) we are concerned with avoiding most.

The relative preference for avoiding the two types of error will depend on how the dichotomised biomarker is to be used. For example, if a safe and inexpensive vaccine were available for AD at the pre-clinical stage, we would likely wish to prioritize sensitivity over specificity when screening patients from an apparently healthy population. Conversely, in a clinical trial of a new AD treatment with potentially serious side effects, we would likely prioritize specificity, in order to minimize giving the treatment to truly non-diseased subjects. Such considerations make clear that the “optimal” cut-point c for a particular biomarker will differ depending on the setting in which the dichotomised biomarker is to be used. As we will see, statistical methods for choosing cut-points vary with regards to how explicit the trade-point between sensitivity and specificity is made.

We now describe a number of the most commonly adopted approaches for choosing the cut-point c. For each, we highlight its advantages and disadvantages, and characterise in what sense the estimated cut-point is “optimal”.

Figure 1

Sensitivity (for AD), specificity, and accuracy (assuming 50% prevalence) of total tau, based on data from Mattsson et al.



Controlling sensitivity or specificity

A simple approach is to choose the value of c that results in a test with a particular, researcher specified, value of either sensitivity or specificity. Whether sensitivity or specificity is controlled will depend on the context in which the test is to be used, and in particular which type of error (false positive or false negative) is deemed more important to control. For example Mattsson *et al* recently derived cut-points for CSF biomarkers for clinical AD which gave sensitivity for AD of 85%. For CSF t-tau, this gives a cut-point of $\geq$320ng/L (95% CI 296, 351). Having found the cut-point c which gives 85% sensitivity, one can then estimate the resulting specificity for the dichotomised test. In the data used by Mattsson *et al*, the T-tau cut-point of 320ng/L gives an estimated specificity of 62.5% (Figure 1).

Advantages of this approach are that it is clear and simple, and it is unaffected by the disease prevalence in the population of interest. Furthermore, the approach forces one to consider how the dichotomised biomarker is to be used and which type of error is deemed most important to control. A disadvantage of the approach is that one type of error is controlled for at the expense of completely disregarding the rate of occurrence of the other. Often one will be concerned with controlling both types of errors, but may have a relative preference for avoiding one to the other. This deficiency can be overcome through use of some type of cost function (see 'Minimizing a cost function').

Normal reference limits

An approach sometimes adopted to use a sample of healthy controls to define a 'normal reference limit' (11). This involves estimating a centile of the control distribution below which a large majority of controls lie, with the implication of being above this cut-point being that a subject is 'not normal'. This is equivalent to choosing the cut-point to give a particular specificity. An advantage of such an approach is that only data on controls are necessary in order to estimate the cut-point. However,

only obtaining control data obviously precludes an assessment of the ability of the cut-point to discriminate between cases and controls.

Minimizing a cost function

Arguably the gold standard approach is to estimate the cut-point that maximizes a chosen, so-called, "utility function", or alternatively minimizes a chosen "cost function". For example, we may choose to minimize the cost function:

$$\text{Cost}(c) = RC \times \text{Prevalence} \times (1-\text{Sensitivity}(c)) + (1-\text{Prevalence}) \times (1-\text{Specificity}(c)) \quad [2]$$

where RC denotes the relative cost of a false negative compared to a false positive (12). Here the term "cost" is not necessarily monetary, but refers to our assessment of the overall cost of making the two types of errors. The relative cost will clearly depend on the context in which the test is to be used and what actions will follow from a positive or negative test result. For example, if a positive test will result in further non-invasive and relatively inexpensive confirmatory investigations, while a negative test would result in failure to treat an aggressive disease, one would choose a high value of RC. This would correspond to ensuring the dichotomised test had high sensitivity, at the expense of specificity.

This approach has much to commend itself, since it involves choosing the cut-point that minimizes cost (or maximizes utility) in some general sense. However, the process of quantifying the relative costs of the two types of errors may be difficult, even when the actions and consequences of a positive and negative test result are well defined.

Maximizing accuracy

A commonly adopted approach is to choose the value of c that maximizes the accuracy of the dichotomised test. This can be achieved by calculating the accuracy resulting from each possible choice of cut-point c in a given dataset, and finding the value which gives the greatest accuracy.

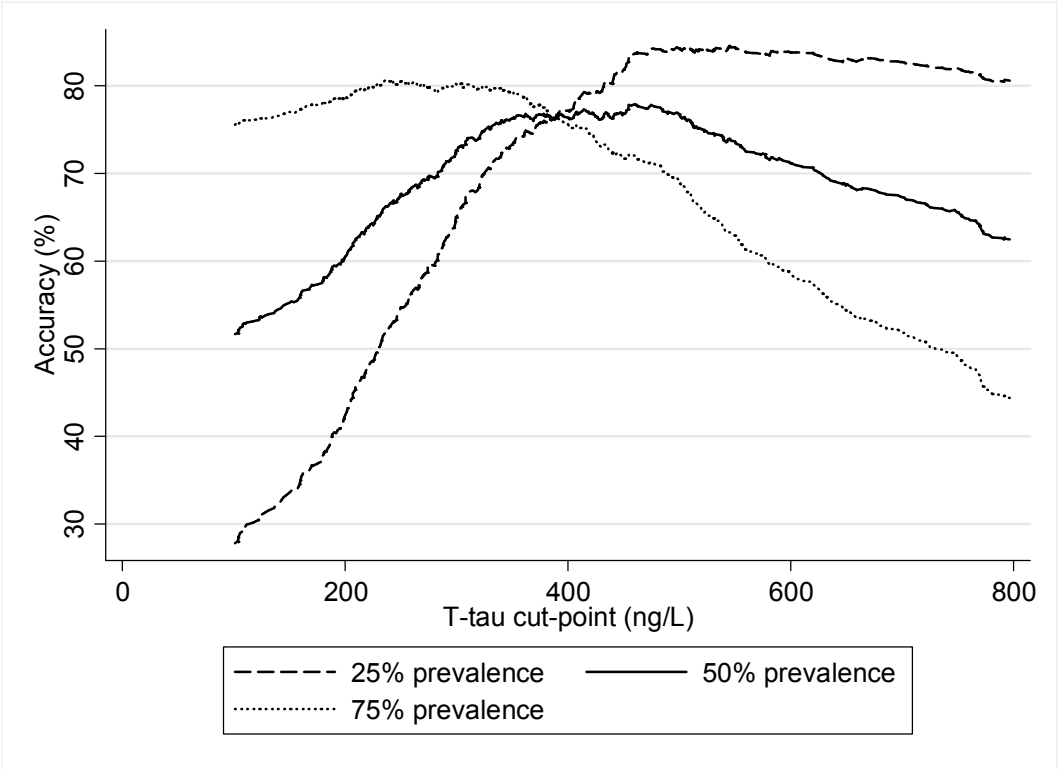
By comparing the formulae for accuracy ([1]) and cost ([2]), we see that maximizing accuracy is equivalent to minimizing cost where we deem a false negative to be equally undesirable as a false positive. Thus in settings where the costs of the two types of errors differ materially, as will often be the case in a clinical setting, choosing the cut-point which maximizes accuracy is inadvisable.

As noted previously, accuracy depends on the prevalence of disease in the target population. This means that the cut-points which maximize accuracy depend on the disease prevalence in the target population. For a very rare disease, accuracy depends almost entirely on the specificity, and so the optimal cut-point will be one for which specificity is increased, at the expense of sensitivity. The reverse holds for a common disease.

This is illustrated in Figure 2, which shows the estimated accuracy of different cut-points for t-tau based on the data of Mattsson et al, using values of disease prevalence of 25%, 50%, and 75%. The optimal cut-point for 25%, 50% and 75% prevalence based on these data can be seen to vary widely at 545ng/L, 460ng/L and 243ng/L respectively.

Figure 2

Estimates of accuracy for T-tau cut-points at 25%, 50% and 75% disease prevalence, based on data from Mattsson et al.



Since accuracy depends strongly on prevalence, if we are to choose the cut-point based on accuracy we must be careful to calculate accuracy values using an estimate of prevalence which is appropriate for the population in which the test is to be used. Cut-points are sometimes derived using data from "case control" studies, where subjects are sampled on the basis of disease status, with accuracy calculated based on the proportion of cases in the sample (2). By virtue of using this sampling method, the proportion of cases in a case-control study will ordinarily not be representative of the population in which the new test is to be used. In this case, the selected cut-point will not be the most accurate in the target population.

Figure 1 shows that there is a wide range of possible cut-points for t-tau that give very similar estimated accuracies yet very different sensitivity/specificity trade-points. This can occur when one (say) decreases the cut-point from a particular value such that sensitivity increases by a similar amount to the decrease in specificity, resulting in accuracy (assuming 50% prevalence) remaining unchanged. A consequence of this is that one may obtain an optimal cut-point that has high sensitivity and low specificity in one dataset, whilst in another dataset a cut-point with identical accuracy can show the reverse.

This point is well illustrated by the CSF cut-points for Aβ1-42 and t-tau derived by Shaw *et al* using autopsy confirmed AD subjects and controls (2). The cut-point maximizing accuracy for t-tau had an estimated sensitivity of 69.6% and specificity of 92.3%, while the cut-point found for Aβ1-42 had a sensitivity of 96.4% and specificity of 76.9%. If the two biomarkers are to be potentially used for the same clinical purpose, considerable caution is required using cut-points which have such radically different sensitivity/specificity trade-points: this is particularly the case if such cut-offs are used to compare the diagnostic properties of different biomarkers (13).

A further consequence of the fact that a large range of cut-points can have similar accuracies is that estimates of the most accurate cut-point may be quite imprecise. This is important, since cut-points are often derived using quite small datasets. The Mattsson *et al* data contain CSF t-tau for 303 controls and 525 patients with AD, which by the standards of CSF biomarker studies is large. Yet the

95% confidence interval for the optimal cut-point based on accuracy (at 50% disease prevalence) is wide, extending from 345ng/L to 479ng/L: this confidence interval is over twice the width of the confidence interval for the cut-point which controls sensitivity at 85%.

One route to mitigating this imprecision is to make some sort of parametric assumption for the biomarker distribution in cases and controls (14). For example, one might assume that the biomarker is normally distributed in cases and controls. Provided such assumptions are reasonable, the estimated optimal cut-point will be more precise. However, violations of the assumptions will result in biased estimates of the optimal cut-point (14). For example, the t-tau data from Mattsson et al have a skewed (hence non-normal) distribution in both controls and AD subjects.

The receiver operating characteristic (ROC) curve

ROC curves plot sensitivity against 1-specificity over the range of possible cut-points c, and are an important tool in diagnostic research (15). ROC curves are often used to show a biomarker's predictive value for a binary outcome (e.g. case/control status), and are sometimes summarised by the area under the ROC curve. The latter has an appealing simple interpretation as the probability that a randomly selected case will have a higher biomarker value than a control (assuming higher values are associated with disease). ROC curves do not in of themselves provide an optimal cut-point. However, a number of the approaches we describe below are based on ROC curves.

Youden's index

For a given cut-point c, Youden's index J(c) is defined as the sum of sensitivity and specificity minus one (16). Finding the cut-point that maximizes Youden's index is equivalent to finding the cut-point that maximizes accuracy for a 50% disease prevalence. The maximum value of J(c) can also be shown to be the maximum vertical distance from the ROC curve to the 45 degree line (corresponding to a biomarker with no predictive value) (17).

One might argue that finding the cut-point that maximizes Youden's index is preferable to maximizing accuracy, since the former does not depend on disease prevalence. An alternative perspective is that it is still equivalent to maximizing accuracy for one, arbitrary, value of disease prevalence. Even if one decides that maximizing accuracy is appropriate, there is no particular reason why the target population in which the biomarker is to be used need have 50% disease prevalence. Since maximizing Youden's index is equivalent to maximizing accuracy for 50% disease prevalence, the former inherits all of the issues of the latter, plus the fact that disease prevalence has effectively been arbitrarily chosen at a level which in general will differ from that in the target population.

Closest to the corner of the ROC curve

An approach sometimes used is to find the cut-point whose point on the ROC curve lies closest to the top-left corner (0,1). The rationale for this is that the ROC curve for a perfect biomarker extends to the top-left corner, and this method thus determines the cut-point that is closest to "perfection" (12). However, it has been shown that this approach is equivalent to minimizing a function of sensitivity and specificity ($(1\text{-sensitivity})^2+(1\text{-specificity})^2$), which has no reasonable justification (12).

Controlling predictive values or risk

Particularly when D represents whether subjects will develop the disease of interest in the future, a popular approach to cut-point selection is to choose c such that subjects with $X \geq c$ have risk of developing disease above a certain threshold p. This is equivalent to choosing c such that the PPV $P(D=1|X \geq c)$ is greater than the desired risk threshold p. Estimating a cut-point in this way thus requires an estimate of disease prevalence and a choice of risk threshold. The latter will naturally depend on the context and the actions following positive and negative test results.

A sophistication of this approach first models risk of disease as a continuous function of the biomarker and then chooses c such that the risk of disease at the cut-point c is equal to a chosen level. Subjects with a positive test result are those who have risk of disease greater than or equal to this risk cut-off. Estimating two risks, one in those who have values above (or equal) to the cut-off and one in those with values below the cut-off can be done empirically from the data at hand without the need for a statistical model. In contrast estimating risk as a continuous function of the biomarker requires assumptions about the data to be made. Two different, but related, approaches involve the use of logistic regression and discriminant function analysis. Royston and Thompson explain the relationship between these two approaches, as well as describing how these techniques have been used in developing screening tests for Down's Syndrome (18). We will not consider such model based approaches further in this paper, other than to note that they are particularly useful when one has results from a number of different biomarkers that one wishes to combine in order to predict those at high risk of a disease. A single risk-based criteria for a positive test (like that for the screening test for Down's syndrome) is almost certain to be superior to an ad-hoc rule based on a combination of different biomarker cut-points.

Mixture modelling

When selecting cut-points for CSF biomarkers in AD, disease classification is often performed on the basis of clinical diagnosis rather than autopsy confirmed disease status. Since even in the best centres a clinical diagnosis of AD is proven incorrect in a proportion of individuals at autopsy, and as it is clear that the earliest pathological changes of AD occur many years prior to clinical symptoms, classification of subjects on the basis of clinical diagnosis is inevitably associated with a degree of mis-classification if the goal is to diagnose pathological AD. Such mis-classification would adversely affect estimates of sensitivity and specificity, and may also bias estimates of optimal cut-points.

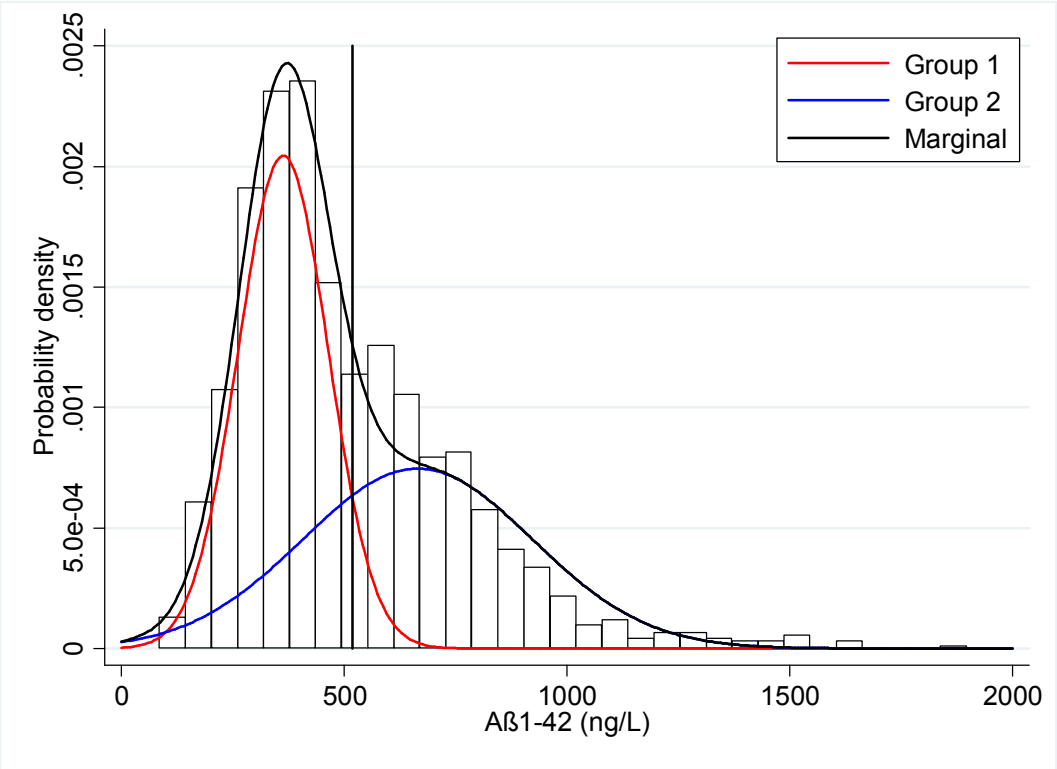
De Meyer *et al* (19) fitted normal mixture models to CSF biomarker data from control, MCI and AD subjects from the Alzheimer's disease Neuroimaging Initiative (ADNI) using data from all three groups, ignoring group membership. The mixture model was fitted with two components, on the assumption that subjects belong to one of two groups (subjects with AD and healthy controls). They then derived a cut-point for CSF Aβ1-42 as the value where the estimated normal component distributions intersected when assigning equal weight to the two.

Following De Meyer et al, we fitted a 2-component normal mixture model to the CSF Aβ1-42 data of Mattsson et al, using data from all control, MCI and AD subjects. Figure 3 shows a histogram of the observed values, the estimated normal components based on fitting a 2-component normal mixture model to the data, and the resulting marginal distribution based on the estimated mixing proportions of 51.8% (left hand component) and 48.2% (right hand component). For these data there is considerable overlap in the estimated normal components. The scaled distributions intersect at an Aβ1-42 level of 518 ng/L. Assuming the two identified normal components correspond to marker values in disease and non-diseased subjects, choosing this intersection corresponds to maximizing accuracy at a disease prevalence of 51.8% (since low Aβ1-42 is associated with AD). If we were to instead use a 50/50 mixing proportion, corresponding to maximizing accuracy for 50% prevalence (or maximizing Youden's index (17)), the intersection occurs at 514 ng/L. In cases where the estimated mixing proportions differ materially from 50% one would obtain larger differences in cut-points between these two approaches (assuming equal weights or using the estimated mixing proportions).

Figure 3

Histogram of CSF Aβ1-42 using data (from Mattsson et al) from controls, MCIs and ADs. Density functions shown in red and blue correspond to estimated distributions from a 2-component normal

mixture model and the resulting marginal distribution is shown in black. The (scaled) normal components intersect at 518 ng/L (indicated by vertical black line).



This mixture modelling approach assumes that the combined sample (i.e. ignoring clinical group membership) consists of subjects who belong to one of two distinct sub-populations, one of which contains subjects with AD pathology and the other without. In reality in a heterogeneous group comprising elderly normal controls, patients with MCI and AD, subjects are likely to lie on a spectrum in terms of AD pathology. It is then unclear exactly what the identified components truly represent, which De Meyer et al themselves acknowledged ("the identified mixture components had no meaning as such"). Indeed, part of De Meyer et al's "validation" of the derived cut-points involved estimating the mixing proportions of the two identified components separately according to clinical diagnosis (control, MCI, AD), thus seemingly relying on the potentially mis-classified disease status they originally sought to avoid using. De Meyer et al therefore also partially validated the cut-points using two external datasets (one with autopsy confirmation of AD, the other consisting of MCI subjects who developed AD within approximately 5 years). Only extended follow-up, ultimately to post-mortem, allows the performance of such dichotomized biomarkers (as predictors of future pathological AD confirmation) to be assessed.

Lastly, we note that since choosing a cut-point at the intersection of two sub-distributions is equivalent to choosing the cut-point which maximizes accuracy (either for 50% prevalence or for the prevalence corresponding to the estimated mixing proportions), the former approach inherits the aforementioned issues of the latter.

Other issues

Quantifying the uncertainty in the optimal cut-point
Whichever approach is adopted to estimate the optimal cut-point it is highly advisable, yet uncommon in published research, to report not only the (estimated) optimal cut-point, but a measure of the uncertainty in the estimate. This enables an assessment of how confident one can be that the value of c found is truly optimal (under the chosen metric) in the population at large. The

standard approach to this is to calculate a confidence interval (typically 95%) for the optimal cut-point c. This interval gives a range of cut-points for which the data are consistent with being optimal. The paper by Mattsson *et al* and our analyses of their data show that confidence intervals for optimal cut-points can be quite wide even when they are estimated from moderately large datasets.

Confidence intervals for the optimal cut-point can easily be found when the cut-point is chosen to constrain sensitivity or specificity at a particular value, using standard statistical package commands for estimating centiles. Finding confidence intervals for cut-points that maximize accuracy is less straightforward. For our analysis of the data from Mattsson *et al* we used a bootstrap resampling approach to find a 95% confidence interval, which required only a small amount of programming in the Stata statistical package. Providing researchers with simple software applications to calculate confidence intervals for cut-points is clearly an area for future work.

Sample sizes
Ideally sample sizes for studies used to estimate cut-points should be chosen so that the cut-point can be estimated with adequate precision. How this calculation is performed will depend on the statistical approach chosen to estimate the cut-point. As for quantification of uncertainty in the estimated cut-point, as far as we are aware little research has been conducted on sample size estimation for cut-point studies, and so this is an important area for future work.

Using the same data to choose the cut-point and estimate performance
Often studies use the same dataset to estimate a cut-point and to assess the diagnostic performance of the biomarker dichotomised at that cut-point. Depending on the approach used to select the cut-point, this can result in over-estimates of the diagnostic utility of the dichotomised biomarker, particularly in small studies (20). Statistical techniques such as cross-validation can be used to allow for the fact that the same data have been used for cut-point selection and assessment of diagnostic utility (20).

Standardisation and inter-lab variability
In this review we have restricted ourselves to the statistical issues related to cut-point determination. Critical to the utility of any diagnostic biomarker is standardisation of biosample collection, minimizing differences between assays, and demonstrating reproducibility and stability of a given assay over time. These issues are of obvious importance when a cut-point derived in a particular study is to be applied in different centres/laboratories, and indicate a need for global harmonisation and standardisation of biomarker measurements.

Conclusions

In this review we have described some of the most commonly used statistical approaches for cut-point selection and some of the issues involved in selecting a cut-point. Critically, the process of choosing a cut-point must be made in light of the way in which the dichotomised biomarker is to be used, the decisions which will follow from positive and negative test results, and the relative costs of making false positive and false negative errors. Furthermore, from a purely statistical perspective, there is no such thing as a unique optimal cut-point, since optimality depends on the metric used to quantify the dichotomised biomarker's performance. From this it follows that there is no single statistical approach which will always be the most appropriate to use.

Future Perspective

We have identified a number of potential pitfalls and highlighted neglected issues. We believe it is inappropriate to estimate cut-points that maximize accuracy based on arbitrary estimates of disease

prevalence in a case/control study. Cut-points are often estimated using relatively small datasets, yet are usually reported without any measure of their estimation uncertainty. Our analysis of the CSF data from Mattsson *et al* illustrates that confidence intervals for optimal cut-points can be wide even when derived from moderately large datasets, and we would encourage these to be reported in future studies. A specific issue in the context of biomarkers for AD is the relative difficulty of obtaining definitive diagnosis of pathological disease, and clearly for the purposes of early accurate diagnosis and drug discovery it will continue to remain important to confirm the diagnosis at autopsy. Furthermore, while dichotomisation has clear advantages in terms of simplification, it may be well be more appropriate to characterise intermediate biomarker levels or "grey zones", resulting in a test that is more complex to interpret, but ultimately is more informative. Operationalizing the new criteria for clinical AD will need to take these issues into account

Finally, it is important to remember that any cut-point derived by any statistical method is irrelevant if the consistency by which the biomarker is measured cannot be assured. This involves strict acceptance criteria for new batches of analytical reagents, the use of longitudinal internal control samples covering clinically relevant biomarker concentrations to demonstrate that the analytical method is stable over time in relation to the study in which the cut-point was determined, and ongoing efforts to harmonise and standardise the collection and measurement of biomarkers across multiple centres.

Executive summary

Biomarkers in Alzheimer's disease

- Biomarkers are playing an increasingly important role in the diagnosis of Alzheimer's disease, in therapeutic trials and observational studies
- In a number of settings it may be advantageous to dichotomise continuous biomarkers as positive or negative, according to whether their value lies above or below a "cut-point".

Applications of dichotomised biomarkers

- Dichotomised biomarkers have potential applications in a diverse range of applications, e.g. in screening at risk populations, in aiding clinical diagnoses, and as part of inclusion criteria for clinical trials.
- Choosing a cut-point at which to dichotomise a biomarker involves making a trade-off between sensitivity to detect disease, and specificity.
- How this trade-off is made must be made in reference to the context in which the dichotomised biomarker is to be used and what actions will follow from a positive and negative test result.

Statistical properties of a dichotomised test

- Sensitivity and specificity quantify, respectively, what proportion of diseased and non-diseased subjects are given a positive and negative test result.
- Positive and negative values quantify the probability of disease given a subject's test result. These quantities depend on the disease prevalence in the target population.
- The accuracy of a test is the proportion of subjects who are correctly classified. Accuracy also depends on disease prevalence.

Statistical methods for choosing cut-points

- There are a range of different statistical methods for estimating cut-points.

- Different methods make different assumptions and find optimal cut-points according to different criteria. Consequently the “optimal” cut-point depends on the assumptions made and metric used to measure performance.
- Use of different statistical methods will result in different “optimal” cut-points.
- Since accuracy depends on disease prevalence, cut-points defined on the basis of accuracy depend on prevalence.
- Choosing cut-points to maximize accuracy may result in cut-points with an undesirable sensitivity/specificity trade-off, and wide confidence intervals

Conclusions

- The selection of cut-points for biomarkers must be made in reference to how the dichotomised biomarker is to be used.
- Due to the variety of metrics through which test performance can be measured, from a statistical perspective there is no such thing as a unique optimal cut-point for a given biomarker.

References

1. Andreasen N, Minthon L, Davidsson P *et al*.: Evaluation of CSF-tau and CSF-Abeta42 as diagnostic markers for Alzheimer disease in clinical practice. *Arch Neurol* 58(3), 373-379 (2001).

2. Shaw LM, Vanderstichele H, Knapik-Czajka M *et al*.: Cerebrospinal fluid biomarker signature in Alzheimer’s Disease Neuroimaging Initiative Subjects. *Ann Neurol* 65, 403-413 (2009).
* Original paper which estimated cut-points for CSF biomarkers to maximize accuracy, using data from autopsy confirmed AD subjects and normal controls.

3. Mattsson N, Zetterberg H, Hansson O *et al*.: CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. *JAMA* 302, 385-393 (2009).
* Original paper which estimated cut-points for CSF biomarkers to give 85% sensitivity for AD, using data from AD subjects and controls.

4. Hansson O, Zetterberg H, Buchhave P *et al*.: Association between CSF biomarkers and incipient Alzheimer’s disease in patients with mild cognitive impairment: a follow-up study. *Lancet Neurol* 5(3), 228-234 (2006).

5. Schott JM, Bartlett JW, Fox NC, Barnes J: Increased brain atrophy rates in cognitively normal older adults with low cerebrospinal fluid Aβ1-42. *Ann Neurol*, 68(6):825-34 (2010).

6. Jack CR, Knopman DS, Jagust WJ *et al*.: Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol*, 9(1), 119-128 (2010).

7. Dubois B, Feldman HH, Jacova C *et al*.: Research criteria for the diagnosis of Alzheimer’s disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol* 6(8), 734-746 (2007).

8. Sperling RA, Aisen PS, Beckett LA *et al*. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 7(3), 280-292 (2011).

9. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, (2003).
* Book describing the statistical concepts and methods involved in classification and prediction.

10. Grimes DA, Schulz KF: Uses and abuses of screening tests. *Lancet* 359, 881-884 (2002).
* Original paper highlighting important issues which should be considered when contemplating using a biomarker as a screening test.

11. Sjögren M, Vanderstichele H, Agren H *et al*. Tau and Aβ42 in cerebrospinal fluid from healthy adults 21-93 years of age: establishment of reference values. *Clin Chem* 47(10), 1776-1881 (2001).

12. Perkins NJ, EF Schisterman: The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol*. 163(7), 670-675 (2006).
* Original paper demonstrating that different 'optimal' cut-points can result from the same data through use of statistical methods which quantify test performance differently.

13. Jack CR, Vemuri P, Wiste HJ *et al*.: Evidence for ordering of Alzheimer Disease Biomarkers. *Arch Neurol* 68(12), 1526-1535 (2011).

14. Fluss R, Faraggi D, Reiser B: Estimation of the Youden index and its associated cutpoint point. *Biom J*. 47(4), 458-472 (2005).

15. Krzanowski WJ, Hand DJ. ROC Curves for Continuous Data. CRC Press, 2009.

16. Youden WJ. An index for rating diagnostic tests. *Cancer* 3, 32–5 (1950).

17. Schisterman EF, Perkins NJ, Liu A *et al*.: Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology* 16, 73-81 (2005).

18. Royston P, Thompson SG. Model-based screening by risk with application to Down's syndrome. *Stat Med* 11(2), 257-268 (1992).

19. De Meyer G, Shapiro F, Vanderstichele H *et al*.: Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Arch Neurol* .67, 949-956 (2010).

20. Leeflang MMG, Moons KGM, Reitsma JB *et al*.: Bias in sensitivity and specificity caused by data-driven selection of optimal cutpoint values: mechanisms, magnitude, and solutions. *Clin Chem* 54, 729-737 (2008).