Journal of Pharmaceutical and Biomedical Analysis 55 (2011) 1148–1156

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

A formal comparison of different methods for establishing cut points to distinguish positive and negative samples in immunoassays

Thomas Jaki[a,*], John-Philip Lawo[b], Martin J. Wolfsegger[b], Julia Singer[b], Peter Allacher[b], Frank Horling[b]

[a] Medical and Pharmaceutical Research Unit, Deparment of Mathematics and Statistics, Lancaster University, Lancaster LA14YF, United Kingdom
[b] Baxter Innovations GmbH, Wagramer Strasse 17-19, 1220 Vienna, Austria

ARTICLE INFO

Article history:
Received 9 February 2011
Received in revised form 6 April 2011
Accepted 8 April 2011
Available online 15 April 2011

Keywords:
Anti-drug antibody
Cut point
ELISA
Immunoassay
Immunogenicity

ABSTRACT

Biotechnology derived therapeutics may induce an unwanted immune response leading to the formation of anti-drug antibodies (ADA). As a result the efficacy and safety of the therapeutic protein could be impaired. Neutralizing antibodies may, for example, affect pharmacokinetics of the therapeutic protein or induce autoimmunity. Therefore a drug induced immune response is a major concern and needs to be assessed during drug development. It is therefore crucial to have assays available for the detection and characterization of ADAs. These assays are used to classify samples in positive and negative samples based on a cut point. In this manuscript we investigate the performance of established and newly developed methods to determine a cut point in immunoassays such as ELISA through simulation and analysis of real data. The different methods are found to have different advantages and disadvantages. A robust parametric approach generally resulted in very good results and can be recommended for many situations. The newly introduced method based on mixture models yields similar results to the robust parametric approach but offers some additional flexibility at the expense of higher complexity.

1. Introduction

Biotechnology derived therapeutics may induce an unwanted immune response resulting in the formation of anti-drug antibodies (ADA). As a consequence of the development of ADA efficacy and safety of the therapeutic protein could be impaired. For example, binding or neutralizing antibodies may affect pharmacokinetics or functionality of the therapeutic protein or even induce autoimmunity when the ADA cross-react with endogenous counterparts. In addition, unwanted immune responses may lead to allergic reactions. As a result, drug induced immune responses to a therapeutic protein are a major concern and need to be assessed during drug development.

Consequently there is a need to develop appropriate assays for the detection and characterization of ADA. In 2007, the European Medicines Agency (EMA) published a guideline that describes the general strategy for the development and validation of assays for immunogenicity assessment of biotechnology derived therapeutic proteins [1]. A multi-tiered approach for the testing of patient samples is recommended. In the first instance a screening assay is used for rapid identification of positive samples while subsequently an additional confirmatory assay is used to confirm the results of the screening assay. As a third step, a functional assay for assessment of the neutralizing capacity of antibodies is recommended. Screening, confirmatory and functional assays for detection and characterization of ADA need to be validated [2,3].

A critical step during assay development and validation is the definition of an appropriate cut-off that can be used to distinguish between positive and negative samples in the screening assay. This initial assay needs to be as sensitive as possible to maximize the detection of true positive samples and should be designed to avoid classifying positive samples as negative. A proportion of false positive samples is acceptable as they can be identified by the following confirmatory assay while costs and time urge to take few samples to this second stage. This approach ensures that the assays will detect as many patients who have indeed developed antibodies.

A valid statistical approach needs to be elaborated to define a reliable cut-off value used in screening and confirmatory assays [4]. For defining an appropriate cut point usually control samples obtained from healthy subjects or untreated patients are used. Such a pool of control samples is in most cases of heterogeneous composition, containing sub-populations consisting of true negative samples as well as true and false positive samples. The portion of each sub population has impact on the final cut-off value if one assumes that indeed all samples are truly negative. For example, a high content of true positives in the sample population due to specific pre-existing antibodies used for calculating the cut-off would result in a high number of false negative evaluation of samples.

* Corresponding author.
E-mail address: jaki.thomas@gmail.com (T. Jaki).

0731-7085/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2011.04.006

Consequently it is crucial to use statistical methods that deal with potential (false) positive samples appropriately when determining a cut point. Different strategies to detect and characterize ADA's have been discussed in [4,5] but no formal evaluation of the methods has yet been undertaken.

In this paper we evaluate a variety of established and less established methods for cut point determination. We will introduce the methods in Section 2 before we compare them thoroughly via simulation (Section 3). We conclude with an in-depth discussion and some future directions.

2. Methods to determine cut point

In this section we will describe various methods for determining cut points. Many of the methods are informed by the discussions in [4], although some adjustments have been made to enable automated cut point determination in the simulations to follow. Most importantly no outlier removal is incorporated prior to applying the various methods as different criteria will result in different cut points. Furthermore, the simulated data studied later do not contain outliers and hence such outlier removal will not be necessary. The conclusions made from the evaluation is nevertheless transferable to situations were outliers are present and subsequently removed. Finally note that the methods discussed here establish a fixed cut point. We will briefly highlight how one of the methods can be extended for floating cut points in the data application and the discussion.

2.1. Method 1: 95th percentile

The cut point is found as the 95th percentile of the screening data. This method does not assume a distribution of the measurements and will result in a false positive rate of 5% if indeed all samples are truely negative.

2.2. Method 2: parametric method

The cut-off value is calculated as $\bar{X} + z_{0.95} \times \mathrm{SD}$, where $\bar{X}$ and SD are the mean and standard deviation of the screening measurements respectively and $z_{0.95}$ is the 95% percentile of the standard normal distribution (approximately 1.645). This method assumes that the measurements are normally distributed. If all samples are negative and the normality assumption is satisfied, it will result in a false positive rate of ∼5%.

2.3. Method 3: robust parametric method

The cut point is found as $\tilde{X} + z_{0.95} \times 1.483 \times \mathrm{MAD}$, where $\tilde{X}$ and MAD are the median and median absolute deviation of the screening measurements respectively and $z_{0.95}$ is the 95% percentile of the standard normal distribution as before. This method resembles the parametric method but uses robust estimators of center and spread. It is designed to yield improved results if measurements are not normally distributed and similar results to the parametric method for normal data.

2.4. Method 4: decision tree

A decision tree approach is used to arrive at the cut-off value. The implementation considered here is taken from the left panel of Fig. 1 in [4] and specifically is calculated according to the following steps.

1. Perform a Shapiro–Wilks test [6] to assess normality of the screening data. If the p-value is <0.05 the data are log-transformated.

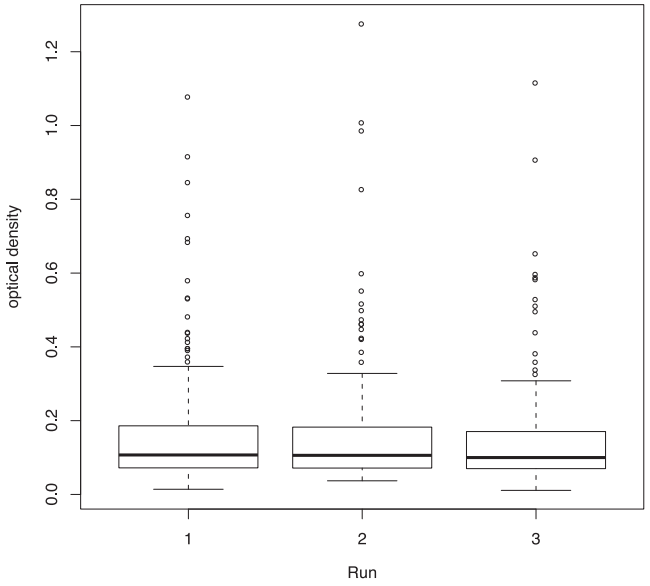

Fig. 1. Boxplot of screening values obtained in three runs of 157 healthy volunteers.

2. Calculate the 25% and 75% percentile, $X_{0.25}$ and $X_{0.75}$, of the (transformed) data. Eliminate all data points outside the interval $[X_{0.25} - 1.5 \times (X_{0.75} - X_{0.25});\ X_{0.75} + 1.5 \times (X_{0.75} - X_{0.25})]$. This corresponds to eliminating data that are classed as outliers in a box-whisker plot (e.g. [7]).
3. Perform the Shapiro–Wilks test [6] to assess normality using the remaining data. If the p-value is <0.05, use the 95% percentile to calculate the intermediate cut point, otherwise the parametric method is used.
4. If data were log-transformed take the anti-logarithm of the intermediate cut point as final cut point otherwise the intermediate cut point is the final cut point.

The above algorithm aims to identify which method is most appropriate by assessing the distribution of the screening values prior to deciding which approach to take. It thereby tries to bring together the advantages of different methods by combining them which comes at the expense that the method used to find the cut point is data dependent and therefore not known a priori.

In general, however, it is not recommended to test every data set for normality, and use the result to decide between parametric and nonparametric statistical tests (e.g. [8–10]). Decisions about when to use parametric or nonparametric tests should be made to cover an entire series of analyses. In addition, with large samples like the ones in immunoassys, minor deviations from normality may be flagged as statistically significant, even though small deviations from a normal distribution will not affect the results.

2.5. Method 5: mixture model

This method, which has not been proposed previously, aims to identify if samples are negative or positive and then only uses the negative samples to find the cut point. It employs so-called (regression) mixture models which have been shown to be useful in many scientific contexts (e.g. [11,12]). A full mathematical description of these models can, for example, be found in [13]. The idea behind such models is that different populations (in this application positive and negative subjects) are described by different probability distributions.

The use of these models here is therefore to firstly identify if there is more than one population in the screening data. If there is more than one population, then only samples belonging to the

Table 1
Cut points and proportion of samples above the cut-off for different methods.

	95th percentile	Parametric method	Robust parametric method	Decision tree	Mixture model
Cut point	0.5108	0.4630	0.2006	0.4395	0.4112
Prop. values above	0.0518	0.0605	0.2224	0.0712	0.0821

larger population, which is assumed to be corresponding to negative samples, will be used for cut point determination while all screening data are used otherwise. Note that this is based on the assumption that the larger component corresponds to negative samples, but alternative selection criteria such as the population with the smaller mean could be used instead. After the appropriate population has been identified, the cut point is found as the 95th percentile here, although any other method could be used at this stage. A formal description and some comments on the specific implementation of this method are found in Appendix A.

2.6. Method 6: mixture model with class predictor

This is a modification to the previous method that allows additional information to be used to find if samples are negative or positive. An additional variable that contains information about the likelihood of a subject being positive (such as a biomarker) is included in the model. In the implementations discussed here the class predictor is derived from using hierarchical clustering on the screening data (details in Appendix A).

2.7. Method 7: experimental approach

The idea of the experimental approach is that one can detect positive values by using screening and confirmatory assay data together. Specifically the cut point is found by:

1. Calculating a preliminary cut point for the confirmatory assay based on the 95% percentile method.
2. Using the preliminary cut point to classify the screening values into positive and negative samples.
3. Creating a new dataset containing all screening samples below the preliminary cut point and all screening samples larger than the preliminary cut-off value provided that the confirmatory value is larger than the screening value. The second set of samples is included as such observations correspond to an unspecific signal.
4. Calculating the final cut point according to the 95% percentile method from the new dataset.

As for the mixture model alternative methods could be used instead of the 95th percentile used here.

3. Comparison of methods

In this section we focus on comparing the different methods to establish cut points introduced in Section 2. We will first illustrate the cut points obtained by the different methods and where they differ on a real data example and then evaluate the methods formally in an extensive simulation study. All analyses and simulations were performed in R [14] Version 2.10.1. The mixture approaches used the `gamlss.mx` package [15] while the package `msm` [16] was used for generating truncated normal distributed data in the simulations.

3.1. Real data examples

To illustrate the various methods and how the cut points differ, we will use the data for one specific protein obtained on 157 healthy volunteers. The data were generated in a direct-binding enzyme-linked immunosorbent assay (ELISA). The microtiter plates (Nunc/Thermo Scientific, Denmark) are coated with a specific protein as antigen and human plasma samples from healthy plasma donors were incubated on the plate. The antigen–antibody complex was detected with a horseradish peroxidase (HRP)-coupled secondary antibody (goat anti-human IgG antibody; AbDSerotec, Germany). The amount of bound secondary antibody was measured by an HRP enzyme-dependent color-change reaction using OPD (o-Phenylenediamine-Dihydrochloride, Sigma Aldrich, Germany) as substrate. The microtiter plates were read with an ELISA reader (Synergy HT; Bio-Tek, USA) in a dual mode at 492 nm measuring wavelength and 630 nm reference wavelength. Delta-OD (=OD at 492 nm minus OD at 630 nm) corrected by the blank value is taken into account as optical density for evaluation.

The data which have been obtained using three runs and two experimenters are illustrated in Fig. 1 and given in full in Table B.4 in Appendix B. From the graph it can easily be seen that a normality assumption is violated due to numerous values outside the box for all runs. Although in practice one would likely try to remove outliers for some methods and transform the data prior to using the parametric method for cut point determination, we will not do so here to highlight the consequence of violating these assumptions on the found cut point.

Table 1 shows the cut-off values found and the proportion of samples above this values for the different methods. Note that the experimental approach is omitted here as no competition values are available. It can be seen that all methods except the robust parametric method yield fairly similar cut points with around 5–8% of the samples exceeding it. The percentile method yields the highest cut point while the robust parametric method obtains the lowest value. The mixture model approach in this instance finds a single log-normal distribution to fit the data best. As there is only one component identified, adding a predictor for class membership will not change the cut-off value either. The most different cut-off value is found by the robust parametric method. Its cut-off value is so much smaller than the other values due to the attempt to correct for the unwarranted assumptions of normality in this case. Implicitly the method weights outliers less which in turn leads to a markedly lower estimate of spread in this instance. Despite this behavior being exaggerated in this example due to omission of transformation and outlier removal, the same behavior is expected regardless.

One feature of the data that has been ignored by the methods discussed above are the different experimenters. The mixture model approach, however, can be adapted to allow for an experimenter specific cut point. In order to obtain such a dynamic cut point, one simply includes the experimenter as a factor in the models fit. The cut point for one experimenter is then simply obtained as before, while the cut point for the second is adjusted by the effect of the experimenter.

In this particular example a single component model fits the data best and the intercept of the model is found as −2.10132 while the coefficient associated with the experimenter is −0.03311. Consequently the cut point for experimenter A is found as the 95th percentile of a log-normal distribution with mean −2.1013 while the cut point for experimenter B stems from a log-normal distribution with mean −2.1344. The variance is the same for both at 0.7503. This results in a cut point of 0.4201 for experimenter A while a cut point of 0.4065 is found for experimenter B. Note that these

cut-off values are in line with the means for each experimenter which at 0.173 is higher for experimenter A than for experimenter B (0.162).

3.2. Simulations

We now turn to evaluating the various methods to establish cut points through simulation which has the advantage that it is exactly known whether a specific value is positive or negative. For a more in-depth evaluation we will consider samples to be either truly positive, false positive or truly negative. For simulation true positive samples have high optical densities in screening assays, but low optical densities in confirmatory assays, false positives have high optical densities in screening and confirmatory assays while true negative samples have low measurements on both assays.

A number of different situations have been studied, but we will focus on the presentation of four illustrative scenarios (additional evaluations are available from the authors upon request). The scenarios differ in the distribution used to generate the optical densities, true positive rates and false positive rates. The true positive rate is the proportion of subjects that have anti-drug antibodies in the population while the false positive rate is the proportion of subject who do not have ADAs despite a high measurement in the screening assay. A basic description for each scenario is given in Table 2, the observable distributions, that is the joint distribution of the samples irrespective of their type, are shown given in Fig. 2 while exact parameters for each scenario are given in Appendix C.

Table 2
Details on generated data.

	Positive vs. negative samples	Distribution	True positive rate	False positive rate
1	No positive samples	Log-normal	0.00	0.00
2	Small difference	Log-normal	0.10	0.10
3	Moderate difference	Normal	0.05	0.05
4	Large difference	Log-normal	0.10	0.05

Note: Normal observations are generated using a truncated normal distribution to ensure positive measurements.

Samples of size 40, 80 and 160 were generated during each of 10,000 simulation runs for each scenario. As sample size had only a limited influence on the performance of each method we will focus attention to the results based on 160 samples unless stated otherwise. To give an impression how varied the cut points found are for the different methods, Fig. 3 shows the distribution of the cut points found for scenario 4.

From the graph it is apparent that the decision tree as well as both mixture model approaches have, sometimes markedly, higher cut points than the other methods. It is also clearly visible that those methods have (substantially) larger variabilities in the found cut points. Furthermore, the variability in the cut points does decrease for all methods as the sample size increases as one would expect. As differences in cut points do not necessarily transfer in better classification, we will now focus on the ability of the different methods to classify samples correctly. To compare the performance of the different methods we will use the false positive rate, false negative

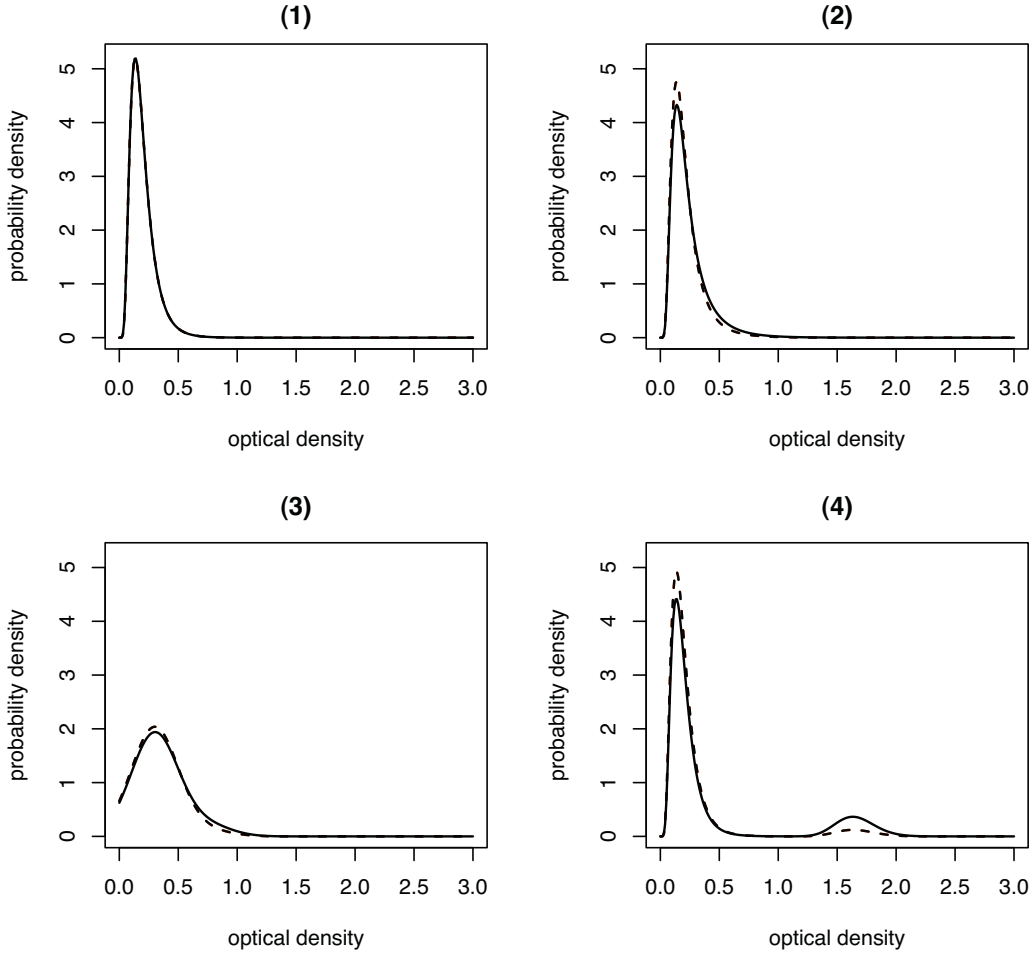

Fig. 2. Observable distributions for simulation scenarios. Solid line corresponds to screening assays while dashed line is for confirmatory assays.

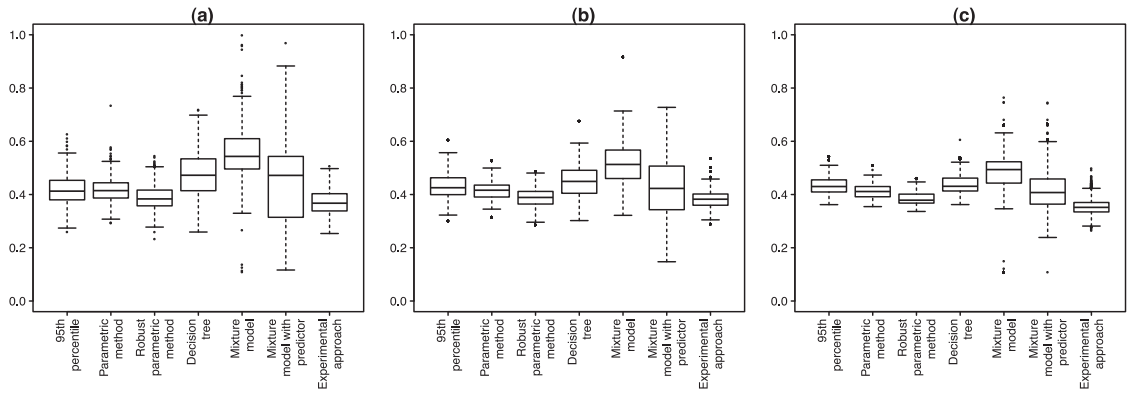

Fig. 3. Distribution of cut points in scenario 4 over 10,000 simulations. (a) corresponds to 40 samples, (b) to 80 samples and (c) to 160 samples.

rate as well as the proportion of correctly classified true positive, true negative and false positive samples. Each measurement will be averaged over the 10,000 simulation runs. Formally the measures are defined as

$$\text{false positive rate} = \frac{\#\,\text{false positive and true negative samples} \geq \text{cut point}}{\text{sample size}}$$

$$\text{false negative rate} = \frac{\#\,\text{true positive samples} \leq \text{cut point}}{\text{sample size}}$$

$$\text{prop. correct true positive} = \frac{\#\,\text{true positive samples} \geq \text{cut point}}{\#\,\text{true positive samples}}$$

$$\text{prop. correct true negative} = \frac{\#\,\text{true negative samples} \leq \text{cut point}}{\#\,\text{true negative samples}}$$

$$\text{prop. correct false positive} = \frac{\#\,\text{false positive samples} \leq \text{cut point}}{\#\,\text{false positive samples}}$$

Ideally a method to determine a cut point would therefore have a false positive and false negative rate close to zero while the proportion of correctly classified samples is close to one. It is worth to note that these measures are solely based on the cut point found and do not take into account of any preliminary classification used when finding the cut-off value. We will begin with an in-depth evaluation of scenario 4 to illustrate which factors appear to be driving the performance of each method for different sample sizes in Table 3.

In the presence of a substantial separation between positive and negative samples and log-normally distributed measurements it is evident that the robust parametric method is superior to the other methods as it classifies all positive samples and almost all negative samples correctly. The only downside of the method is that it also classifies all false positive samples as positive. More generally the overlap between the distributions of positive and negative samples implies that methods determining cut points can only increase the performance in one category at the price of decreasing the ability to correctly identify the other category, a feature clearly seen in the table.

An interesting feature of the mixture approaches is that, although the cut points are highly variable (Fig. 3), the classification results are a close second to the robust parametric method for medium to large sample sizes. This suggests that the variability in the cut points is derived from features in the data which are not picked up by the other methods. One would therefore expect that the mixture approach could be improved even further for specific datasets by allowing mixing of other distributions than ones used in the simulations. Finally it is also evident that the inclusion of a class predictor, which in this context has somewhat arbitrarily

Table 3
Detailed results of classificationfor scenario 4.

n	Method	False positive rate	False negative rate	Correct false positive	Correct true negative	Correct false positive
40	95th percentile	0.0171	0.0637	0.3947	0.9996	0.6129
	Parametric method	0.0429	0.0120	0.9206	0.9999	0.0828
	Robust parametric method	0.0843	0.0000	1.0000	0.9587	0.0000
	Decision tree	0.0418	0.0362	0.7230	0.9880	0.2922
	Mixture model	0.0766	0.0317	0.7628	0.9510	0.2438
	Mixture model with class predictor	0.0668	0.0033	0.9726	0.9765	0.0326
	Experimental approach	0.0963	0.0066	0.9309	0.9405	0.0587
80	95th percentile	0.0155	0.0660	0.3666	1.0000	0.6844
	Parametric method	0.0449	0.0080	0.9422	0.9994	0.0800
	Robust parametric method	0.0850	0.0000	1.0000	0.9586	0.0000
	Decision tree	0.0341	0.0462	0.6023	0.9902	0.4183
	Mixture model	0.0797	0.0288	0.7851	0.9467	0.2448
	Mixture model with class predictor	0.0610	0.0069	0.9444	0.9828	0.0439
	Experimental approach	0.0858	0.0057	0.9408	0.9549	0.0429
160	95th percentile	0.0172	0.0610	0.3654	1.0000	0.6247
	Parametric method	0.0464	0.0036	0.9685	0.9999	0.0334
	Robust parametric method	0.0796	0.0000	1.0000	0.9638	0.0000
	Decision tree	0.0377	0.0368	0.6396	0.9901	0.3584
	Mixture model	0.0738	0.0192	0.8199	0.9586	0.1800
	Mixture model with class predictor	0.0605	0.0013	0.9881	0.9851	0.0114
	Experimental approach	0.0762	0.0038	0.9608	0.9669	0.0295

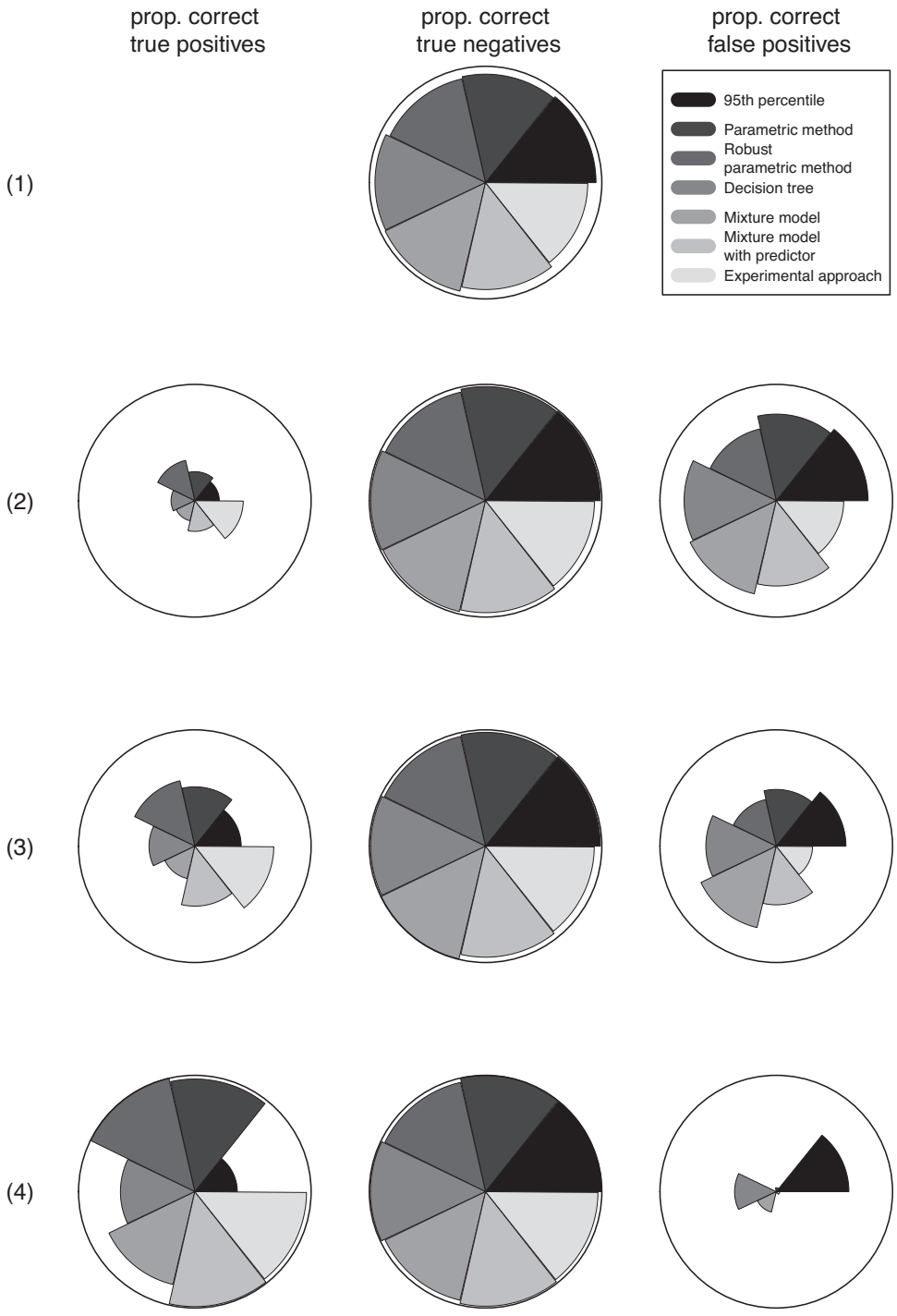

Fig. 4. Comparison of classification rates for seven different methods for cut point determination for four scenarios.

been derived from hierarchical clustering, does improve classification. Using more informative variables as class predictors will likely result in an even larger improvement over the mixture model without class predictor. Unfortunately these improvements do involve tailoring the method(s) for specific applications which makes it not only impossible to evaluate in simulation but also makes the procedure much more time demanding than the algorithmic alternatives.

In Fig. 4 we now compare the different methods amongst themselves across the four scenarios for a sample size of 160. Each row in the graphic corresponds to one scenario, each column to one category and each segment corresponds to one method. The size of the segment corresponds to the proportion of correctly classified samples while the surrounding circle gives the reference of 100% correctly classified samples.

From this comparison it is easy to see that all methods for all scenarios are performing well in identifying negative samples as such with improved classification as separation between positive and negative samples increases. The success in classification of truly positive samples heavily depends on the separation between positive and negative samples as one would expect with classification rates starting around 10% for small separation. Once again the dilemma that a high proportion of correctly identified true positive samples comes at the price of poor identification of false positives is apparent for all methods.

Notable is also that the robust parametric method is clearly superior to the other methods that are based only on screening data when positive samples are present but that the percentile method and the mixture approach are superior when only negative samples are studied. In contrast the decision tree and the parametric method underperform when no positive samples are present and are substantially worse than the best methods in the presence of positive samples. Finally no influence of the distribution (normal or log-normal) on the preformance of the different methods is detectable.

The experimental approach, which in contrast to all other methods uses both screening and confirmatory data, results in superior classification of positive samples for small and moderate separation between positive and negative samples at the price of less successful classification of negative sample which results in a false positive rate of around 8–10%.

Once separation between positive and negative samples is large, however, the robust parametric method and the mixture model approach yield better classification of positive and negative samples.

Overall it therefore appears that the robust parametric method is preferable for samples that include positive samples while the percentile method should be considered when no positive samples are expected. The mixture model approaches perform well in both instances provided enough samples are available and can therefore be recommended if one is uncertain about positive samples being present and particularly if specific tailoring to the application is feasible.

4. Discussion

In this paper seven methods for cut point estimation were formally compared in terms of their ability to identify positive and negative samples. Due to the overlap between the distributions of positive and negative samples, it is inherent to methods determining cut points that the performance to identify a certain category can be increased only at the cost of decreasing the ability to correctly identify the other category, a relationship clearly seen in the simulation results. As these assays are mainly used in drug safety assessments, false positive samples can be regarded as the "company's risk" (a false signal of increased anti-drug antibody level will be obtained), while false negative samples are the "consumers' risk", the truly increased anti-drug antibody level will not be detected potentially resulting in unsafe or uneffective treatment of the patient. Therefore, a conservative approach is to choose a method which ensures a very low level of false negatives. In these terms the robust parametric method is best while the mixture model with class predictor is a close second.

Discussion is focused around the situation in which a large difference between the assay measurements of the positive and negative samples was assumed as differences between methods were expected to be more pronounced in this case although the relative merits of the various methods is largely unchanged by this difference. For all sample sizes, the positive samples were best identified by the robust parametric method with the mixture model as a close second. The highest false negative rates were found with the 95% percentile method, followed by the experimental approach, and the decision tree. 100% of the true positive cases were identified with the robust parametric method followed by the mixed model with class predictor which also correctly identified around 95% of the true positive cases. All methods resulted in a high proportion of correctly identified true negatives.

In this paper we have also discussed two novel approaches focusing for cut point determination namely mixture models and an experimental approach. The underlying idea of both is to identify positive samples and excluding them from the data used to determine the cut-off value. Overall the mixture model approach often yields classification results that are close to best method while the experimental approach gives excellent results for small to medium separation at the price of slightly higher false positive rates. Within these methods the 95th percentile was used on the "clean" dataset although in the light of the results obtained it is advisable to explore the use of the robust parametric method instead. The mixture modelling framework also offers additional flexibility not considered here. Our focus has been on fixed cut points and consequently other factors such as experimenter, day, . . . have not been included when finding cut points. The mixture models presented here can be extended to include such additional variables. These, then called regression mixture models can then be used to find dynamic cut points that are specific to these additional variables.

Moving to the overall aim of the multi-tiered approach for testing a patients sample to identify positive samples, best results are likely obtained by using different methods for each of the stages. Specifically, a method that yields a high number of positive samples at the screening stage while a method identifying a high number of true positives in the confirmatory stage appears promising. Detailed evaluations are planned for the future to examine different combinations of approaches thoroughly.

Conflict of interest

J.-P. Lawo, M.J. Wolfsegger, J. Singer, P. Allacher, F. Horling are employees of Baxter.

Appendix A. Details of cut point determination using mixture models

1. Fit a 1-component mixture model assuming a log-normal distribution to the screening data.
2. Fit a 2-component mixture model assuming a log-normal distribution for one component and a generalized gamma distribution for the other to the screening data.
3. Find the model with the lower Bayesian Information Criteria (BIC, [17]).
   (a) If the 1-component model has been selected, find the 95th percentile of the estimated log-normal distribution.
   (b) If the 2-component model has been selected, find the 95th percentile of distribution corresponding to the larger component.

Notes on the method:

- The generalized gamma distribution is defined as

$$f(y|\mu,\sigma,\nu) = \frac{\theta^{\theta} z^{\theta} \nu e^{-\theta z}}{\Gamma(\theta) y}$$

  where $z = (y/\mu)^{\nu}$, $\theta = 1/(\sigma^2 \times |\nu|^2)$ for $y > 0$, $\mu, \sigma > 0$ and $-\infty > \nu > \infty$;
- Distributional choices have been made based on real data but can be modified for different contexts;
- Other model selection criteria such as AIC could also be used;
- This method assumes that there are more negative than positive subjects in the sample although simple adjustments can be made to relax this assumption;
- When desired, class predictors are included in both models. In the simulations the class predictor has been found by using hierarchical clustering [18] on the screening data and use the class membership derived from the clustering algorithm as class predictor in the model. Complete linkage and the euclidian distance is used in the hierarchical clustering algorithm.

Appendix B. Dataset

Table B.4
Optical densities 157 healthy volunteers for a specific protein using three runs and two experimenters.

Experimenter	A	B	B		A	B	B
Run	1	2	3		1	2	3
1	0.847	1.009	0.654	81	0.045	0.037	0.054
2	0.108	0.108	0.099	82	0.075	0.061	0.066
3	0.185	0.187	0.130	83	0.1870	0.082	0.080
4	0.065	0.081	0.075	84	0.321	0.236	0.304
5	0.107	0.11	0.103	85	0.198	0.176	0.172
6	0.229	0.273	0.216	86	0.055	0.059	0.040
7	0.221	0.153	0.169	87	0.135	0.128	0.100
8	0.081	0.057	0.048	88	0.073	0.071	0.076
9	0.073	0.091	90.0730	89	0.093	0.073	0.084
10	0.095	0.106	0.077	90	0.0950	0.074	0.078
11	0.151	0.178	0.137	91	0.145	0.083	0.119
12	0.325	0.286	0.226	92	0.053	0.039	0.048
13	0.143	0.151	0.127	93	0.091	0.064	0.073
14	0.118	0.140	0.143	94	0.158	0.112	0.141
IB	0.108	0.082	0.071	95	0.065	0.053	0.045
16	0.093	0.116	0.116	96	0.483	0.500	0.440
17	0.081	0.091	0.081	97	0.091	0.084	0.077
18	0.110	0.127	0.096	98	0.581	0.518	0.584
19	0.361	0.3610	0.248	99	0.089	0.068	0.068
20	0.050	0.062	0.044	100	0.079	0.064	0.064
21	0.046	0.053	0.044	101	0.081	0.057	0.073
22	0.308	0.290	0.227	102	0.441	0.286	0.291
23	0.070	0.081	0.060	103	0.111	0.081	0.078
24	0.029	0.105	0.102	104	0.019	0.050	0.064
25	0.160	0.167	0.111	105	0.058	0.059	0.054
26	0.300	0.328	0.269	106	0.283	0.260	0.242
27	0.090	0.085	0.063	107	1.079	1.277	1.117
28	0.439	0.425	0.296	108	0.115	0.090	0.096
29	0.050	0.073	0.063	109	0.129	0.088	0.111
30	0.037	NA	NA	110	0.397	0.234	0.339
31	0.077	0.096	0.087	110.1	0.154	0.139	0.121
32	0.039	0.085	0.075	112	0.131	0.096	0.114
33	0.053	0.061	0.061	113	0.053	0.047	0.059
34	0.048	0.070	0.1060	114	0.088	0.106	0.101
35	0.089	0.112	0.084	115	0.180	0.178	0.154
36	0.076	0.079	NA	116	0.150	0.126	0.121
37	0.176	0.132	0.1320	117	0.158	0.142	0.126
38	0.182	0.211	0.178	118	0.0890	0.089	0.095
39	0.097	0.096	0.089	119	0.058	0.063	0.066
40	0.222	0.198	0.191	120	0.069	0.056	0.071
41	0.039	0.043	0.039	121	0.084	0.065	0.074
42	0.158	0.197	0.133	122	0.19	0.091	0.097
43	0.347	0.312	0.259	123	0.685	0.553	0.589
44	0.112	0.130	0.13	124	0.534	0.463	0.512
45	0.027	0.056	NA	125	0.07	0.075	0.070
46	0.048	0.051	0.038	126	0.113	0.114	NA
47	0.160	0.134	0.080	127	0.051	0.055	0.046
48	0.092	0.124	0.118	128	0.265	0.293	0.258
49	0.067	0.080	0.066	129	0.917	0.987	0.908
50	0.067	0.084	0.069	130	0.168	0.141	0.127
51	0.186	0.244	0.166	131	0.288	0.210	0.212
52	0.041	0.050	NA	132	0.264	0.158	0.228
53	0.145	0.159	0.156	133	0.120	0.117	0.111
54	0.085	0.126	0.0560	134	0.061	0.056	0.065
55	0.172	0.198	0.196	135	0.397	0.387	0.308
56	0.758	0.828	0.598	136	0.414	0.449	0.360
57	0.054	0.047	0.011	137	0.240	0.298	0.287
58	0.119	0.140	0.123	138	0.095	0.080	0.094
59	0.063	0.073	0.061	139	0.322	0.283	0.306
60	0.103	0.112	0.104	140	0.0720	0.072	0.070
61	0.107	0.134	0.103	141	0.097	0.078	0.086
62	0.063	0.088	0.068	142	0.235	0.156	0.187
63	0.113	0.073	0.081	143	0.084	0.071	0.087
64	0.086	0.117	0.109	144	0.0	0.094	0.090
65	0.424	0.109	0.089	145	0.532	0.600	0.530
66	0.014	NA	NA	146	0.220	0.207	0.162
67	0.333	0.294	0.179	147	0.190	0.231	0.153
68	0.069	0.071	0.058	148	0.107	0.1	0.101
69	0.091	0.085	0.078	149	0.059	0.049	0.054
70	0.093	0.101	0.0	150	0.158	0.098	0.112
71	0.064	0.066	0.062	151	0.105	0.085	0.105
72	0.374	0.475	0.383	152	0.138	0.083	0.088

Table B.4 (*Continued*)

Experimenter	A	B	B		A	B	B
Run	1	2	3		1	2	3
73	0.036	0.041	0.037	153	0.072	0.066	0.071
74	0.055	0.056	0.057	154	0.322	0.277	0.215
75	0.060	0.064	0.063	155	0.060	0.054	0.055
76	0.162	0.166	0.139	156	0.077	0.070	0.059
77	0.392	0.422	0.327	157	0.120	0.126	0.119
78	0.125	0.134	0.172	Mean	0.173	0.169	0.155
79	0.084	0.067	0.074				
80	0.695	0.463	0.497				

Mean experimenter A: 0.173.
Mean experimenter B: 0.162.

Table C.5
Parameters used to generate data for simulations.

Scenario	Distribution	Mean negative	Mean positive	Std deviation negative	Std deviation positive	True positive rate	False positive rate
1	Log-normal	−1.75	NA	0.5	NA	0.00	0.00
2	Log-normal	−1.75	50.0	0.5	0.5	0.05	0.05
3	Normal	0.30	0.75	0.2	0.2	0.05	0.05
4	Log-normal	−1.75	0.5	0.5	0.1	0.10	0.05

Appendix C. Supplementary information

See Table C.5.

References

[1] Committee for medicinal products for human use, Guideline on immunogenicity assessment of biotechnology derived therapeutic proteins, European Medicines Agency (2007). URL: http://www.ema.europa.eu/docs/enGB/documentlibrary/Scientificguideline/2009/0 (last visited 29.10.2010).

[2] A. Mire-Sluis, Y. Barrett, V. Devanarayan, E. Koren, H. Liu, M. Maia, T. Parish, G. Scott, G. Shankar, E. Shores, S. Swanson, G. Taniguchi, D. Wierda, L. Zucker-man, Recommendations for the design and optimization of immunoassays used in the detection of host antibodies against biotechnology products, J. Immunol. Methods 289 (2004) 1–16.

[3] S. Gupta, S. Indelicato, V. Jethwa, T. Kawabata, M. Kelley, A. Mire-Sluis, S. Richards, B. Rup, E. Shores, S. Swanson, E. Wakshull, Recommendations for the design, optimization, and qualification of cell-based assays used for the detection of neutralizing antibody responses elicited to biological therapeutics, J. Immunol. Methods 321 (2007) 1–18.

[4] G Shankar, V. Devanarayan, L. Amaravadi, Y. Barrett, R. Bowsher, D. Finco-Kent, M. Fiscella, B. Gorovits, S. Kirschner, M. Moxness, T. Parish, V. Quarmby, H. Smith, W. Smith, L. Zuckerman, E. Koren, Recommendations for the validation of immunoassays used for detection of host antibodies against biotechnology products, J. Pharma. Biomed. Anal. 48 (2008) 1267–1281.

[5] E. Koren, H. Smith, E. Shores, G. Shankar, D. Finco-Kent, B. Rup, Y.-C. Barrett, V. Devanarayan, B. Gorovits, S. Gupta, T. Parish, V. Quarmby, M. Moxness, S. Swanson, G. Taniguchi, L. Zuckerman, C. Stebbins, A. Mire-Sluis, Recommendations on risk-based strategies for detection and characterization of antibodies against biotechnology products, J. Immunol. Methods 333 (2008) 1–9.

[6] S. Shapiro, M. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (1965) 591–611.

[7] J. Chambers, W. Cleveland, B. Kleiner, P. Tukey, Graphical Methods for Data Analysis, Wadsworth & Brooks/Cole, 1983.

[8] S. Bailey, Subchronic toxicity studies, in: S. Chow, J. Liu (Eds.), Design and Analysis of Animal Studies in Pharmaceutical Development, Marcel Dekker, New York, 1998, pp. 135–196.

[9] C. Hennig, Falsification of propensity models by statistical tests and the goodness-of-fit paradox, Philos. Math. 15 (2007) 166–192.

[10] M. Wolfsegger, T. Jaki, B. Dietrich, J. Kunzler, K. Barker, A note on statistical analysis of organ weights in non-clinical toxicological studies, Toxicol. Appl. Pharmacol. 240 (2009) 117–122.

[11] M. Wedel, W. Desarbo, A review of recent developments in latent class regression models, in: R. Bagozzi (Ed.), Advanced Methods of Marketing Research, Blackwell Publishers, Cambridge, 1994, pp. 352–388.

[12] M.L. Van Horn, T. Jaki, S.L. Ramey, K. Masyn, J.A. Smith, S. Antaramian, Assessing differential effects: Applying regression mixture models to identify variations in the influence of family resources on academic achievement, Dev. Psychol. 45 (2009) 1298–1313.

[13] G. McLachlan, D. Peel, Finite Mixture Models, John Wiley & Sons, Inc., New York, 2000.

[14] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0, URL: http://www. R-project.org.

[15] M. Stasinopoulos, B. Rigby, gamlss.mx: a GAMLSS add on package for fitting mixture distributions, r package version 4.0-4, 2010, URL: http://CRAN.R-project.org/package=gamlss.mx.

[16] C. Jackson, msm: multi-state Markov and hidden Markov models in continuous time, r package version 0.9.7, 2010, URL: http://CRAN.R-project.org/package=msm.

[17] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.

[18] J. Ward, Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963) 236–244.