Title: A false sense of security? Can tiered approach be trusted to accurately classify immunogenicity samples?

Author: Thomas Jaki Peter Allacher Frank Horling

PII: S0731-7085(16)30270-9
DOI: http://dx.doi.org/doi:10.1016/j.jpba.2016.05.031
Reference: PBA 10674

To appear in: *Journal of Pharmaceutical and Biomedical Analysis*

Received date: 19-2-2016
Revised date: 18-5-2016
Accepted date: 19-5-2016

Please cite this article as: Thomas Jaki, Peter Allacher, Frank Horling, A false sense of security? Can tiered approach be trusted to accurately classify immunogenicity samples?, <*![CDATA[Journal of Pharmaceutical and Biomedical Analysis]]*> (2016), http://dx.doi.org/10.1016/j.jpba.2016.05.031

A false sense of security? Can tiered approach be trusted to accurately classify immunogenicity samples?

Thomas Jaki[a,*], Peter Allacher[b,c], Frank Horling[b]

[a] *Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom*
[b] *Baxalta Innovations GmbH, Industriestrasse 67, 1221 Vienna, Austria*
[c] *IMC University of Applied Sciences, Piaristengasse 1, 3500 Krems an der Donau, Austria (currently)*

Abstract

Detecting and characterizing of anti-drug antibodies (ADA) against a protein therapeutic are crucially important to monitor the unwanted immune response. Usually a multi-tiered approach that initially rapidly screens for positive samples that are subsequently confirmed in a separate assay is employed for testing of patient samples for ADA activity. In this manuscript we evaluate the ability of different methods used to classify subject with screening and competition based confirmatory assays. We find that for the overall performance of the multi-stage process the method used for confirmation is most important where a t-test is best when differences are moderate to large. Moreover we find that, when differences between positive and negative samples are not sufficiently large, using a competition based confirmation step does yield poor classification of positive samples.

Keywords: Anti-drug antibody, confirmatory, cut point, immunoassay, immunogenicity, screening, specificity

1. Introduction

Detecting and characterizing of anti-drug antibodies (ADA) against a protein therapeutic are crucially important to monitor the unwanted immune response. Usually a multi-tiered approach that initially rapidly screens for positive samples that are subsequently confirmed in a separate assay is employed for testing of patient samples for ADA

[*] Corresponding author. Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom. E-mail address: jaki.thomas@gmail.com, Phone: +44 (0) 1524 59 23 18

presence. Several regulatory guidelines [1, 2, 3] and white papers [4, 5, 6] describe the testing strategies, assay formats, validation requirements and performance expectations for such assays have been published.

In order to use either screening or confirmatory assays, establishing cut points that are used to classify into negative and positive samples are paramount. An upper negative limit of 95% for the screening cut point is recommended [1, 2, 4, 6], resulting in a 5% false-positive rate. The subsequent confirmation assay used here aims to eliminate false positive samples based on competition assays. These competition assays are a tool to identify possible signal contribution from unspecific antibody binding and additionally analyze all samples using a study-drug inhibited assay. This assay is basically set up identically to the uninhibited assay with the exception that all samples are pre-incubated with excess amount of free specific protein antigen ("antigen competition"). Specific antibodies directed against the particular antigen are bound in the form of immune complexes in the liquid phase and subsequently removed during washing steps. Hence, the specificity of antibodies detected with the uninhibited assay can be confirmed by a reduction of signal in the inhibited assay. Recently various methods for finding cut points for screening assays [7, 8] and confirmatory assays [9] have been evaluated.

One of the unexpected and striking findings when evaluating the performance of confirmatory assays [9] was that extremely large differences between uninhibited and inhibited samples are necessary to separate positive from negative samples. This surprising finding led us to investigate the capability of the multi-tier approach to separate positive and negative samples. In this manuscript we will evaluate the ability of the multi-tier approach for classifying samples in both simulations and real data evaluations.

2. Classifying samples

Previously a large number of different approaches for classifying screening [e.g. 6, 7] and confirmatory assays [e.g. 9] have been described. In this evaluation we consider 7 methods to be used in screening assays and three approaches for confirmatory assays yielding 21 different combination of approaches. We have attempted to be as comprehensive as possible in the methods investigated, yet the sheer number of approaches currently in the literature disallowed a full evaluation. The most notable ideas that have not been considered here is the simplified decision tree in [10] and the fixed percent inhibition method [6, 9]. The former was excluded as initial evaluations revealed an undistinguishable performance to the decision tree in [6] while the latters subjective choice of what percentage ought to be used was prohibitive.

In this section we will describe the different methods for classifying samples. The principle idea of each approach for confirmatory assays is to determine if the change in assay signal with and without pre-incubation of a sample with high amounts of the therapeutic drug is large enough to be a relevant indicator to distinguish between true positive and false positive samples. We will therefore consider the situation where measurements without pre-incubation for each sample are available (the screening data) and that measurements with and without preincubation are available for confirmation. For the latter we also assume pre-incubation is successful and truly leads to inhibition. Moreover, we assume that multiple runs (analyses) per sample are undertaken and that measurements are corrected for run noise. As in [9] we will use an average of the runs per sample (e.g. mean per subject across runs) to utilize multiple runs recognizing that more involved methods may be necessary depending on the underlying experimental design [e.g. 11]. Measurements with pre-incubation of the therapeutic drug will be referred to as “inhibited measurements” and without incubation as “uninhibited measurements”.

2.1. Methods for classification: Screening assays

Method S1: 95th percentile

The cut point is found as the 95th percentile of the uninhibited observations.

Method S2: Parametric method

The cut-off value is calculated as $\bar{X}$+$\mathrm{z}_{0.95}$*SD, where $\bar{X}$ and SD are the mean and standard deviation of the uninhibited measurements respectively and $\mathrm{z}_{0.95}$ is the 95% percentile of the standard normal distribution (approximately 1.645).

Method S3: Robust parametric method

The cut point is found as $\tilde{X}$+$z_{0.95}$*1.483*MAD, where $\tilde{X}$ and MAD are the median and median absolute deviation of the uninhibited measurements respectively and $z_{0.95}$ is the 95% percentile of the standard normal distribution as before.

Method S4: Decision tree

The following decision tree, as described in [6], is used to find the cut-point.

1. Perform a Shapiro-Wilks test [12] to assess normality of the uninhibited data. If the p-value is < 0.05 the data are log-transformated.
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*($X_{0.75}$ - $X_{0.25}$); $X_{0.75}$ + 1.5*($X_{0.75}$ - $X_{0.25}$)]. This corresponds to eliminating data that are classed as outliers in a box-whisker plot [e.g. 13].
3. Perform the Shapiro-Wilks test [12] 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.

Note, that in general 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. 14, 15]. This procedure has, however, been proposed as a compromise between statistical rigour and practicality.

Method S5: Mixture model

This method, which has been proposed in [7], aims to identify if samples are negative or positive and then only uses the negative samples to find the cut point. The approach uses (regression) mixture models [e.g. 16, 17, 18] that allow different populations (in this application postive and negative subjects) to follow different probability distributions.

The approach is to firstly identify, using the Bayesian Information Criterion (BIC) if there is more than one population in the screening data. If there is more than one population, then only samples belonging to the 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. The cut point is then found as the 95th percentile of the observations. A formal description and details on the specific implementation of this method are provided in the supplementary materials.

Method S6: Prediction intervals

This approach is advocated in [8] and is based on obtaining intervals for future observations based on m historical observations. In particular the cut-point is found as $\bar{X}+\mathrm{t}_{0.95,m-1}$*SD*$\sqrt{1+1/m}$, where $\bar{X}$ and SD are the mean and standard deviation of the uninhibited measurements respectively and $\mathrm{t}_{0.95,m-1}$ is the 95% percentile of a t-distribution with $m-1$ degrees of freedom.

Method S7: Experimental approach

The experimental approach, which utilizes screening and confirmatory assay data together obtains the cut point through the following steps:

1. Find a preliminary cut point for the inhibited samples based on the 95% percentile method;
2. Use the preliminary cut point to classify uninhibited values into positive and negative samples;
3. Create 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 nonspecific signal (false positives);

4. Use the 95% percentile method with the new dataset to get the final cut-point.

2.2. Methods for classification: Confirmatory assays

Method C1: Parametric difference

Find the difference between uninhibited and inhibited measurement for each sample

$$D = \text{uninhibited measurement} - \text{inhibited measurement}.$$

The cut point is found as $c_D = \bar{D} + z_{0.999} * \sigma_D$ where $\bar{D}$ is the average difference across all samples, σ_D is the corresponding standard deviation and $z_{0.999}$ is the 99.9% percentile of the standard normal distribution (approximately 3.09).

Method C2: Parametric % inhibition

For each sample find the percent change in inhibition as

$$\mathrm{I} = 100 * \left(1 - \frac{\text{inhibited measurement}}{\text{uninhibited measurement}}\right)$$

The inhibition based cut point is found as $c_I = \bar{I} + z_{0.999} * \sigma_I$ where $\bar{I}$ is the average percent change in inhibition across all samples, σ_I is the corresponding standard deviation and $z_{0.999}$ is the 99.9% percentile of the standard normal distribution as before.

Method C3: t-test

Perform a one-sided 2-sample t-test of all runs of the log-transformed study drug inhibited values against the log-transformed uninhibited values for each sample. If the resulting p-value is less than 0.01 the sample is classed positive.

3. Simulation of multi-tiered approach

We begin by considering simulations of the 2-stage classification approach in this section. This has the advantage that it is exactly known whether a specific value is positive or negative, allowing for an informed comparison of the different approaches. 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 show high measurements when uninhibited, but low values under inhibition, false positives have high measurements when uninhibited and inhibited while true negative samples have low measurements under both conditions.

We will generate data for this evaluation in two parts. In the first part, data that are used to determine the cut-points are simulated from a population that only contains negative samples. Both inhibited and uninhibited samples will be generated and we will use 160 samples in the first part of the evaluation as previous work [7, 9] suggests limited impact of sample size. The second part of the data are used to evaluate the performance of the classification methods and cut-points found based on the first set of data. The data are generated to contain 85% true negative samples, 10% of the data are truly positive and 5% are false positive samples. To ensure accurate estimation of the classification rates we will simulate 1,000 samples and estimate the classification rates based on these data. Both normal and log-normally distributions are evaluated and 1,000 simulation runs are performed. Three runs will be used for establishing cut points and evaluating classification. Table 1 in the supplementary materials shows the exact parameters used to generate the data. Note that, while only a limited set of evaluations are presented here, many more simulations have been run. As the conclusions from these were qualitatively the same as the once presented, we have omitted them here for brevity.

To evaluate the performance of the classifications we will look at the proportion of correctly classified true positive, true negative and false positive samples averaged over 1,000 simulation runs. We begin, however, by considering the number of samples that are selected for confirmation as this number has direct implications for the practicability of the classification method. Note, that we expect around 200 observations to be classed as positive at this stage, as 10% of the 1000 observations are truly positive, 5% are false positive and the cut-points are found so that 5% error in classing negative samples are allowed. Figure 1 shows the distribution of the number of samples that are classed as positive based on the screening data for the seven different methods. The first notable observation is that the experimental approach classes almost twice as many observations as positive than the other approaches. Consequently the risk of missing a positive signal at this stage is lower for that approach while at the same time the risk of including large numbers of truly negative samples in the confirmation step is also increased. Secondly, the difference in number of samples classed positive is (on average) quite similar for all the other approaches although more variability is observed in the mixture approach. It is however notable that for the situation with a small difference between positive and negative samples, only about 100 samples are considered positive and hence a high risk of false negatives exists, while the larger differences between positive and negative samples yield numbers quite close to the expected 200 samples. Additional evaluations (not shown) suggest, that the number of positive samples is very stable once the difference between positive and negative samples is sufficiently large. For normally distributed data, for example, this difference needs to be around 2.5 standard deviations.

$\sim\sim$ Figure 1 about here $\sim\sim$

Next we evaluate the ability of the various approaches to classify correctly. The objective of this evaluation is two-fold. Firstly we wish to see how well commonly used classification approaches for immunogenicity assays work in realistic situations and secondly determine which approach (that is which combination of methods for cutpoint calculation for screening assay and confirmation assay) is best. We begin by focusing on the overall classification rates, when the robust parametric approach, which in [7] is found to be one of the best performing methods, is used for the screening assays. Figure 2 shows a clear separation between the methods for classification for confirmatory assays investigated. The % inhibition methods performs far worse than the other two approaches in classifying true positive samples, when the robust parametric method is used for the screening assays. The difference between the parametric difference and the t-test is more nuanced, however. The t-test performs best classifying true positive samples - only for large difference between positive and negative samples the parametric difference is marginally better. When looking at the classification performance of the different approaches of samples that are truly negative, the parametric difference is slightly better, although the t-test also results in a large proportion of correct classifications. This small difference is expected, however, as the number of correct classifications is dominated by the method used for the screening data. The parametric difference is clearly superior to the t-test in classifying false positives which only achieves about 80% correct classifications. It also shows a peculiar dip in the proportion of correct classifications for medium differences between positive and negative samples.

∼∼ Figure 2 about here ∼∼

The evaluation shown in Figure 2 focuses on the situation, where the robust parametric difference is used for the screening assays. Although the classification rates do differ slightly, when using other methods during the screening phase, the relative patterns described above are the same. It is notable, that the difference between positive and negative samples needs to be quite large in order to see good classification of true positive samples, while the classification of true negatives and false positives is much less effected by that difference. To investigate the combination of methods further, we now look at the different methods for classifying screening assay data. Figure 3 show the overall classification rates for each screening classification method when the confirmation uses the t-test. The parametric and the robust parametric method result in the best classification rates for truly positive samples while all methods appear to give good classification of negative samples. The mixture model approach and the prediction interval are best in determining false positive samples. Overall, the percentile approach and the parametric methods appear to provide best results. As the difference between methods is most pronounced for small differences between the positive and negative samples, the graph displays a 1 and 1.2 standard deviation difference for normal and log-normal data, respectively. The overall patterns are the same as this difference increases, however. It is worth noting that the methods become almost indistinguishable for differences that classify an adequate proportion of subjects correctly.

∼∼ Figure 3 about here ∼∼

The evaluations so far clearly indicate that the method used for screening has little bearing on the overall ability to classify correctly when using a 2-tier approach while the method used for confirmation is of high importance. A simple t-test performs best in classifying true positive samples but does not do so well in classifying false positives. In contrast a simple difference approach has good classification for false positives, yet only results in adequate classification rates if the difference between positive and negative samples are very large. This raises the immediate question whether a confirmatory assay should be used at all. To investigate this further we contrast the classification rates after screening only and after screening and confirmation. The robust parametric method is used for the screening assays while the parametric difference is used for confirmation.

∼∼ Figure 4 about here ∼∼

Figure 4 shows that using a confirmatory assay has an notable effect on the ability to classify positive sample correctly for small to moderate differences between positive and negative samples. At the same time the confirmatory assays do result in a much improved false positive rate. When using the % inhibition approach the results are even worse in terms of classifying positive samples. The results for the t-test are closer to the ones obtained by using screening alone but result in much worse classification for false positive samples (see Figure 2).

4. Differences in methods for a specific example

The previous evaluations were based on simulated data but suggest that it may not be beneficial for classifying ADA positive and ADA negative samples to use a confirmation assay. In this section we consider a real dataset (illustrated in Figure 5 and full dataset provided in Table 2 of the supplementary materials) to highlight where the different approaches lead to distinct conclusions.

The data set was generated by means of a direct-binding enzyme-linked immunosorbent assay (ELISA). The ELISA was designed to detect total Ig antibodies (i.e. isotypes IgG, IgM and IgA) specifically directed against a particular protein antigen. Plasma samples from 160 clinically healthy plasma donors were analyzed, each using three runs with and without inhibition. For the uninhibited runs, micro-titer plates (Nunc/ThermoScientific, Denmark) were coated with the particular protein antigen. Human plasma samples from healthy plasma donors (Baxter AG, Austria) were incubated on the plate at a dilution of 1:20. Antibodies directed against the antigen that were present in the samples bound to the antigen. After several washing steps, the antigen-antibody complexes was detected using a horseradish peroxidase (HRP)-coupled secondary antibody (goat anti-human Ig antibody; AbD Serotec, Germany). The amount of bound secondary antibody was measured by an HRP enzyme-dependent color-change reaction using TMB (3,3',5,5'- tetramethylbenzidine solution, AbD Serotec, Germany) as substrate. The color reaction is directly proportional to the amount of bound antibodies. The micro-titer plates were subsequently read with a plate photometer (ELISA reader Synergy HT; Bio-Tek, USA) in a dual mode at 450nm measuring wavelength and 630nm reference wavelength. The dual mode allows the elimination of measurement errors due to scratches or dirt on the micro-titer plates. Delta-OD (=optical density at 450nm minus optical density at 630nm) corrected by the blank value is taken into account as optical density (OD) for evaluation. Each sample was analyzed in independent triplicates by two different analysts on different days.

As a tool to identify possible signal contribution from unspecific antibody binding, all samples were additionally analyzed using a study-drug inhibited assay (i.e. confirmatory assay). This assay is basically set up identically to the uninhibited assay with the exception that all samples are pre-incubated with excess amount of free specific protein antigen (antigen competition). Specific antibodies directed against the particular antigen are bound in the form of immune complexes in the liquid phase and subsequently removed during washing steps. Hence, the specificity of antibodies detected with the uninhibited assay can be confirmed by a reduction of OD signal in the inhibited assay.

Close examination of the data shows that a large variability both between subjects and between runs exists in this dataset. Similarly one can observe that the responses increase after the addition of the antigen for a number of subjects. This is somewhat surprising as it is not consistent with the inhibition model and suggests some other confounding factor. In such a situation a more advanced modeling approach that accounts for this confounding factor may be called for. For the purpose of illustration, however, we will keep with the basic approaches and illustrate the difference in final classification resulting from different combinations of methods.

∼∼ Figure 5 about here ∼∼

Figure 6 shows the number of samples that are classed as positive for the different stages and methods. It can be seen that the different methods for screening classify between 8 and 19 observations as positive. Looking at the methods for confirmation within each of the screening results it is firstly notable that the inhibition based method classes the most samples as positive while the difference based method does not class any as positive. More interestingly, however, is the fact that, despite substantially different numbers of samples being classed possitive during screening, the confirmation step does yield very consistent results. The t-test and the difference based method class exactly one and none sample, respectively, as positive, irrespective of the screening method used. This underlines once more how large the impact of the confirmation step is in comparison to the screening step.

∼∼ Figure 6 about here ∼∼

5. Discussion

In this paper we have evaluated the ability of the multi-tier approach to classify positive and negative samples. We find that, irrespective of the specific methods used for determining cutpoints for screening and competition assay the approach is able to correctly identify truly negative samples as such. Similarly there is high confidence in the correct classification of false positive samples. Unfortunately, however, we also find that in general the two-tier approach only identifies positive samples correctly if very large differences between positive and negative samples are present. For small differences between positive and negative samples positive samples are frequently misclassified. We also find that this performance at small differences between inhibited and uninhibited samples is due to a lack of sensitivity of the methods of classification for the competition based confirmatory assay used in this study. As a consequence, samples with a low signal in the screening assay should not be applied to the competition based confirmatory assay because of the low confidence of a correct true positive evaluation. Instead, a lower limit for confirmed positive samples should be introduced in addition to the lower limit of detection of any antibody in the screening assay [20]. For moderate and small differences using the competition based confirmatory approach decreases the number of correctly classified positive samples drastically.

In our evaluation we have focused on simple methods for classification (a summary of the preformance for all combinations of methods is given in Table 3 of the supplementary materials) and have not considered more complex methods such as [11]. We have done so as the simulated conditions we have considered meant that these simple approaches were appropriate. It is clear, however, that more complex real life settings and experimental designs will require more complex methods for analysis. Similarly we have focused on scenarios that did not provide any particular additional challenges such as positive samples when establishing the screening cut point. It is clear that the findings still have general applicability even if more challenging scenarios are considered.

Acknowledgements

This report is independent research arising in part from Prof Jaki's Career Development Fellowship and Senior Research Fellowship (NIHR-CDF-2010-03-32, NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

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/en_GB/document_library/Scientific_guideline/2009/09/WC500003947.pdf last visited 2013-02-14.

[2] Center for Drug Evaluation and Research, Guidance for Industry. Assay development for immunogenicity testing of therapeutic proteins. Draft guidance., U.S. Department of Health and Human Services Food and Drug Administration. (2009). URL http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM192750.pdf last visited 2013-02-14

[3] United States Pharmacopeia Convention, United States Pharmacopeia and National Formulary (USP 35-NF 30), United States Pharmacopeia Convention, 2012.

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

[5] E. Koren, H. W. 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. J. Swanson, G. Taniguchi, L. A. Zuckerman, C. C. Stebbins, A. R. Mire-Sluis, Recommendations on risk-based strategies for detection and characterization of antibodies against biotechnology products, J Immunol Methods 333 (1-2) (2008) 1–9.

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

[7] T. Jaki, J.-P. Lawo, M. J. Wolfsegger, J. Singer, P. Allacher, F. Horling, A formal comparison of different methods for establishing cut points to distinguish positive and negative samples in immunoassays, J Pharm Biomed Anal 55 (5) (2011) 1148–1156.

[8] D. Hoffman, M. Berger, Statistical considerations for calculation of immunogenicity screening assay cut points, J Immunol Methods 373 (1-2) (2011) 200–208.

[9] T. Jaki, J.-P. Lawo, M. J. Wolfsegger, P. Allacher, F. Horling, A comparison of methods for classifying samples as truly specific with confirmatory immunoassays, J Pharm Biomed Anal 88 (2014) 27–35.

[10] L. Zhang, J. J. Zhang, R. J. Kubiak, H. Yang, Statistical methods and tool for cut point analysis in immunogenicity assays, J Immunol Methods 389 (2013) 79–87.

[11] F. Schaarschmidt, M Hofmann, T Jaki, B Grün, L. A. Hothorn, Statistical approaches for the determination of cut points in anti-drug antibody bioassays, J Immunol Methoods 418 (2015) 84-100.

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

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

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

[15] M. J. Wolfsegger, T. Jaki, B. Dietrich, J. A. Kunzler, K. Barker, A note on statistical analysis of organ weights in non-clinical toxicological studies, Toxicol Appl Pharmacol 240 (1) (2009) 117–122.

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

[17] L. Van Horn, T. Jaki, S. Ramey, K. Masyn, J. Smith, S. Antaramian, L. A, Assessing differential effects: Applying regression mixture models to identify variations in the influence of family resources on academic achievement, Developmental Psychology 45 (5) (2009) 1298–1313.

[18] M. L. Van Horn, T. Jaki, K. Masyn,G. Howe, D. J. Feaster, A. E. Lamont, M. R. W. George, M. Kim, Evaluating differential effects using regression interactions and regression mixture models, Educational and Psychological Measurement 75 4 (2015) 677-714.

[19] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (2) (1978) 461–464.

[20] S. F. J. Whelan, C. J. Hofbauer, F. M. Horling, P. Allacher, M. J. Wolfsegger, J. Oldenburg, C. Male, J. Windyga, A. Tiede, H. P. Schwarz, F. Scheiflinger, B. M. Reipert, Distinct characteristics of antibody responses against factor VIII in healthy individuals and in different cohorts of hemophilia A patients. Blood 121 6 (2013) 1039–1048.

Figure 1: Boxplots of the number of samples exceeding the screening cut-point for the different methods. Panels (a) and (b) display normally distributed data with 1 and 5 standard deviation difference between positive and negative samples, respectively. Panels (c) and (d) display log-normally distributed data with 1.2 and 4 standard deviations difference between positive and negative samples, respectively.

Figure 2: Classification rates across the two stages when the robust parametrics method is used for the screening assays and different approaches are utilized for the confirmation. A range of differences between positive and negative samples is investigated.

Figure 3: Classification rates across the two stages when the t-test is used for the confirmatory assays and different approaches are utilized for the screening assays. A difference of 1 and 1.2 standard deviations between positive and negative samples for normal and log-normal data, respectively are used.

Figure 4: Classification rates when the robust parametric method is used for screening and a parametric difference is used for the confirmatory assays. A range of differences between positive and negative samples is investigated.

Figure 5: Histogram of the average screening and competition values per subject.

Figure 6: Histogram of number of samples classified as positive after screening (big boxes) and after confirmation for the different methods.

Table 1: Parameters used to generate data for simulations. Between run correlation was 0.7 and correlation between uninhibited and inhibited samples 0.3.

		Uninhibited samples		Inhibited samples				prop	prop
Distribution	Stage	μ_n	$\mu_{tp} = \mu_{fp}$	μ_n	μ_{tp}	μ_{fp}	σ	tp	fp
Normal	1	0.3	NA	0.2	NA	NA	0.2	0.00	0.00
	2	0.3	0.5, 0.6, . . . 2.0	0.2	0.2	0.4, 0.5, . . . 1.9	0.2	0.10	0.05
Log-Normal	1	0.3	NA	0.2	NA	NA	0.2	0.00	0.00
	2	0.3	0.5, 0.6, . . . 2.0	0.2	0.2	0.4, 0.5, . . . 1.9	0.2	0.10	0.05

μ is mean and σ is standard deviation parameters of distribution. n. . . negative, tp . . . true positive, fp . . . false positive, prop tp . . . proportion of truly positive samples per stage, prop fp . . . proportion of false positive samples per stage

Table 2: Measurements with and with-out study drug inhibition of 160 healthy volunteers with 3 runs each used in the example provided.

	uninhibited samples			inhibited samples				uninhibited samples			inhibited samples		
#	run 1	run 2	run 3	run 1	run 2	run 3	#	run 1	run 2	run 3	run 1	run 2	run 3
1	0.284	0.212	0.200	0.310	0.270	0.278	81	0.132	0.131	0.183	0.188	0.246	0.208
2	0.215	0.302	0.149	0.225	0.318	0.278	82	0.162	0.186	0.202	0.239	0.308	0.363
3	0.085	0.138	0.094	0.103	0.137	0.114	83	0.119	0.127	0.151	0.194	0.219	0.241
4	0.161	0.218	0.135	0.153	0.243	0.154	84	0.080	0.092	0.093	0.133	0.180	0.179
5	0.219	0.290	0.185	0.276	0.334	0.270	85	0.103	0.119	0.126	0.116	0.160	0.184
6	0.353	0.463	0.205	0.244	0.393	0.279	86	0.127	0.150	0.137	0.187	0.231	0.223
7	0.185	0.243	0.158	0.211	0.321	0.311	87	0.136	0.113	0.151	0.149	0.140	0.163
8	0.155	0.309	0.111	0.206	0.341	0.142	88	0.090	0.126	0.115	0.164	0.245	0.234
9	0.093	0.089	0.097	0.217	0.192	0.145	89	0.232	0.460	0.261	0.259	0.620	0.317
10	0.079	0.152	0.084	0.133	0.178	0.165	90	0.096	0.122	0.143	0.110	0.193	0.137
11	0.166	0.245	0.127	0.153	0.222	0.247	91	0.220	0.245	0.278	0.296	0.275	0.341
12	0.062	0.118	0.065	0.110	0.160	0.161	92	0.206	0.268	0.286	0.229	0.261	0.350
13	0.148	0.210	0.113	0.205	0.186	0.178	93	0.120	0.122	0.143	0.224	0.290	0.247
14	0.137	0.161	0.091	0.176	0.177	0.175	94	0.421	0.566	0.508	0.444	0.534	0.414
15	0.165	0.223	0.131	0.209	0.251	0.179	95	0.100	0.123	0.128	0.188	0.228	0.194
16	0.245	0.551	0.133	0.345	0.786	0.136	96	0.178	0.129	0.165	0.212	0.192	0.179
17	0.164	0.156	0.118	0.178	0.198	0.114	97	0.110	0.160	0.146	0.163	0.308	0.216
18	0.207	0.387	0.218	0.207	0.273	0.277	98	0.093	0.135	0.133	0.151	0.224	0.259
19	0.147	0.257	0.153	0.222	0.309	0.247	99	0.233	0.313	0.341	0.277	0.365	0.448
20	0.151	0.274	0.133	0.165	0.209	0.166	100	0.077	0.105	0.117	0.175	0.193	0.199
21	0.091	0.163	0.080	0.172	0.286	0.192	101	0.541	0.600	0.497	0.258	0.333	0.354
22	0.075	0.178	0.069	0.101	0.227	0.114	102	0.220	0.262	0.200	0.274	0.309	0.278
23	0.143	0.216	0.110	0.227	0.298	0.159	103	0.177	0.225	0.265	0.261	0.325	0.504
24	0.101	0.352	0.077	0.218	0.519	0.149	104	0.101	0.157	0.087	0.146	0.219	0.160
25	0.092	0.091	0.085	0.176	0.162	0.129	105	0.196	0.259	0.260	0.196	0.319	0.263
26	0.075	0.100	0.121	0.157	0.155	0.159	106	0.115	0.169	0.183	0.179	0.231	0.227
27	0.057	0.082	0.062	0.130	0.179	0.137	107	0.268	0.415	0.475	0.279	0.391	0.423
28	0.126	0.117	0.116	0.189	0.232	0.221	108	0.750	0.936	1.005	0.711	0.899	1.155
29	0.116	0.155	0.089	0.204	0.269	0.169	109	0.448	0.693	0.515	0.303	0.438	0.417
30	0.117	0.147	0.097	0.158	0.186	0.103	110	0.190	0.180	0.224	0.249	0.311	0.362
31	0.255	0.377	0.111	0.242	0.398	0.114	111	0.115	0.116	0.126	0.155	0.181	0.174
32	0.074	0.150	0.085	0.140	0.204	0.136	112	0.209	0.238	0.267	0.243	0.290	0.253
33	0.146	0.202	0.017	0.192	0.305	0.178	113	0.133	0.167	0.148	0.201	0.372	0.302
34	0.177	0.246	0.257	0.242	0.293	0.338	114	0.177	0.204	0.231	0.223	0.282	0.354
35	0.167	0.212	0.200	0.203	0.240	0.279	115	0.234	0.306	0.201	0.272	0.352	0.317
36	0.086	0.100	0.123	0.144	0.147	0.165	116	0.199	0.214	0.213	0.236	0.273	0.349
37	0.990	1.212	1.066	0.310	0.351	0.371	117	0.422	0.560	0.482	0.314	0.436	0.470
38	0.099	0.108	0.073	0.190	0.249	0.223	118	0.116	0.165	0.149	0.175	0.271	0.235
39	0.224	0.298	0.213	0.258	0.294	0.265	119	0.172	0.208	0.202	0.226	0.336	0.289
40	0.100	0.184	0.087	0.118	0.212	0.107	120	0.157	0.168	0.143	0.219	0.211	0.228
41	0.159	0.250	0.236	0.334	0.344	0.427	121	0.135	0.311	0.189	0.140	0.332	0.196
42	0.139	0.148	0.211	0.193	0.264	0.316	122	0.160	0.233	0.274	0.259	0.387	0.341
43	0.069	0.063	0.096	0.110	0.121	0.151	123	0.079	0.091	0.113	0.136	0.209	0.193
44	0.074	0.085	0.094	0.121	0.140	0.144	124	0.092	0.125	0.133	0.182	0.258	0.237
45	0.282	0.319	0.304	0.392	0.432	0.516	125	0.217	0.250	0.252	0.286	0.374	0.330
46	0.135	0.195	0.186	0.291	0.438	0.475	126	0.068	0.070	0.077	0.133	0.183	0.161
47	0.167	0.240	0.201	0.278	0.402	0.342	127	0.099	0.092	0.111	0.170	0.223	0.138
48	0.140	0.278	0.151	0.219	0.366	0.285	128	0.197	0.138	0.238	0.298	0.284	0.282
49	0.132	0.183	0.145	0.184	0.251	0.194	129	0.166	0.134	0.079	0.202	0.173	0.183
50	0.156	0.169	0.215	0.149	0.179	0.222	130	0.114	0.081	0.161	0.188	0.219	0.273
51	0.073	0.080	0.099	0.124	0.171	0.149	131	0.137	0.129	0.189	0.207	0.225	0.294
52	0.080	0.120	0.124	0.162	0.239	0.179	132	0.094	0.086	0.106	0.133	0.159	0.178
53	0.135	0.121	0.216	0.164	0.286	0.288	133	0.264	0.180	0.302	0.321	0.345	0.434
54	0.102	0.118	0.126	0.146	0.200	0.174	134	0.175	0.148	0.156	0.216	0.253	0.257
55	0.265	0.295	0.260	0.251	0.289	0.275	135	0.171	0.170	0.246	0.198	0.243	0.315
56	0.115	0.233	0.228	0.208	0.293	0.190	136	0.183	0.200	0.166	0.212	0.294	0.269
57	0.124	0.172	0.172	0.213	0.227	0.193	137	0.274	0.281	0.187	0.277	0.378	0.292
58	0.080	0.097	0.140	0.155	0.201	0.215	138	0.323	0.341	0.437	0.286	0.247	0.364
59	0.186	0.217	0.247	0.286	0.457	0.363	139	0.195	0.205	0.181	0.196	0.242	0.264
60	0.058	0.079	0.093	0.094	0.140	0.086	140	0.198	0.171	0.244	0.179	0.185	0.238
61	0.144	0.192	0.159	0.171	0.222	0.212	141	0.165	0.157	0.175	0.226	0.246	0.344
62	0.111	0.140	0.148	0.213	0.304	0.267	142	0.099	0.128	0.093	0.125	0.168	0.170
63	0.173	0.194	0.195	0.163	0.257	0.176	143	0.187	0.230	0.260	0.197	0.209	0.258

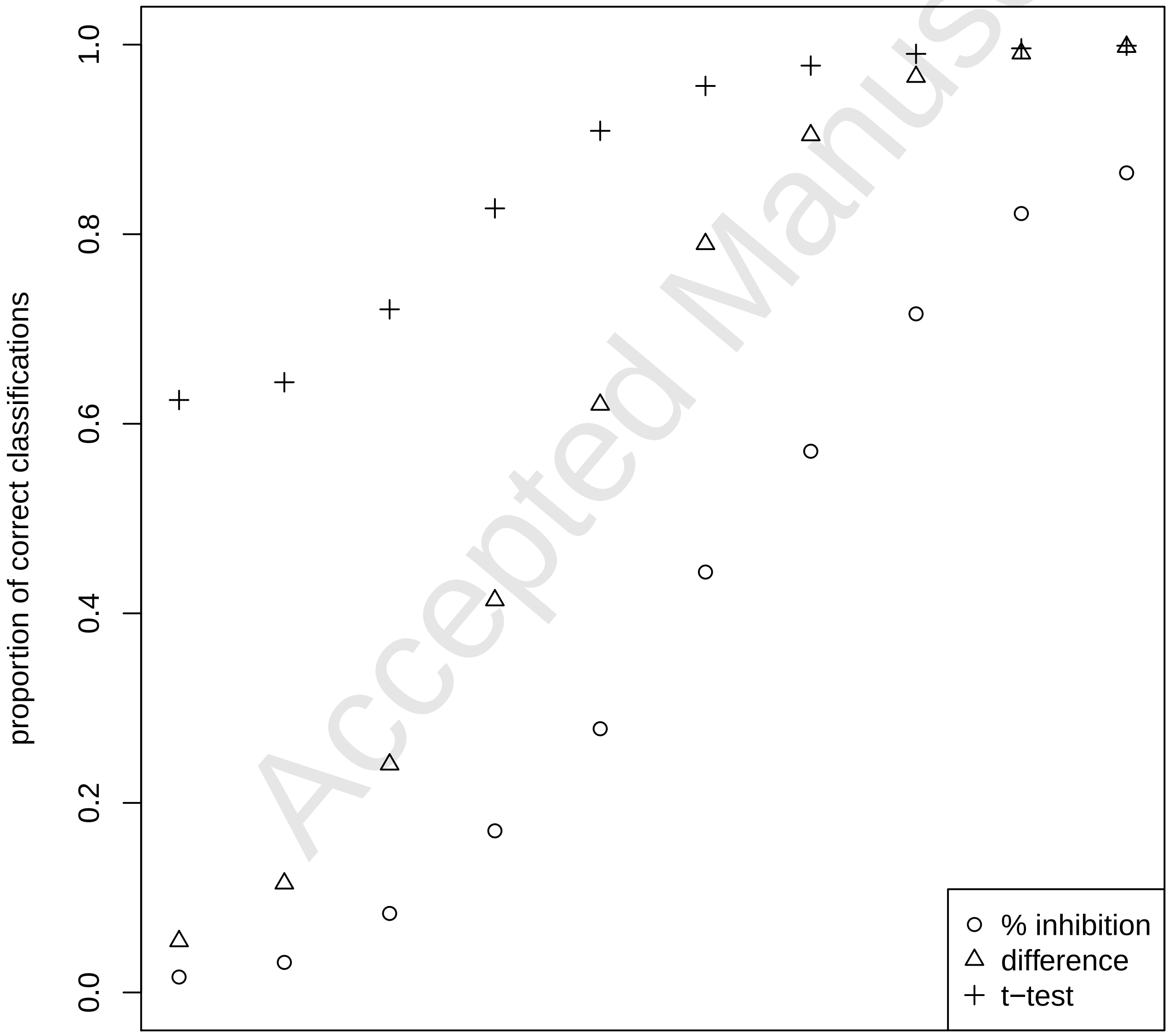
Table 3: Summary conclusions for performance of multi-tier approach across all combinations of methods.

Method		Overall
Screening	Confirmation	performance
95th percentile	Parametric difference	Good
Parametric method		Good
Robust parametric method		Moderate
Decision tree		Poor
Mixture model		Poor
Prediction intervals		Poor
Experimental approach		Poor
95th percentile	Parametric % inhibition	Poor
Parametric method		Poor
Robust parametric method		Poor
Decision tree		Poor
Mixture model		Poor
Prediction intervals		Poor
Experimental approach		Poor
95th percentile	t-test	Good
Parametric method		Good
Robust parametric method		Moderate
Decision tree		Moderate
Mixture model		Moderate
Prediction intervals		Moderate
Experimental approach		Poor

415

Highlights

- The multi-tier approach for classification of immunoassays is evaluated
- The methods are illustrated on a real dataset
- The methods are compared via simulation
- We find that the overall performance of the multi-stage process is dominated by the method used for confirmation



Mean difference between screening and confirmatory assay