Journal of Immunological Methods 418 (2015) 84–100

Contents lists available at ScienceDirect

Journal of Immunological Methods

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

Computational modeling

Statistical approaches for the determination of cut points in anti-drug antibody bioassays

Frank Schaarschmidt [a,*], Matthias Hofmann [b], Thomas Jaki [c], Bettina Grün [d], Ludwig A. Hothorn [a]

[a] Institute of Biostatistics, Leibniz Universität Hannover, Herrenhaeuser Str. 2, D-30419 Hannover, Germany
[b] Novartis Institutes for Biomedical Research, CH-4057 Basel, Switzerland
[c] Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK
[d] Department of Applied Statistics, Johannes Kepler University Linz, Austria

ARTICLE INFO

Article history:
Received 31 July 2014
Received in revised form 12 February 2015
Accepted 12 February 2015
Available online 27 February 2015

Keywords:
Anti-drug antibody
Immunoassays
Immunogenicity
Screening cut point
Software

ABSTRACT

Cut points in immunogenicity assays are used to classify future specimens into anti-drug antibody (ADA) positive or negative. To determine a cut point during pre-study validation, drug-naive specimens are often analyzed on multiple microtiter plates taking sources of future variability into account, such as runs, days, analysts, gender, drug-spiked and the biological variability of unspiked specimens themselves. Five phenomena may complicate the statistical cut point estimation: i) drug-naive specimens may contain already ADA-positives or lead to signals that erroneously appear to be ADA-positive, ii) mean differences between plates may remain after normalization of observations by negative control means, iii) experimental designs may contain several factors in a crossed or hierarchical structure, iv) low sample sizes in such complex designs lead to low power for pre-tests on distribution, outliers and variance structure, and v) the choice between normal and log-normal distribution has a serious impact on the cut point.
We discuss statistical approaches to account for these complex data: i) mixture models, which can be used to analyze sets of specimens containing an unknown, possibly larger proportion of ADA-positive specimens, ii) random effects models, followed by the estimation of prediction intervals, which provide cut points while accounting for several factors, and iii) diagnostic plots, which allow the post hoc assessment of model assumptions. All methods discussed are available in the corresponding R add-on package mixADA.

1. Introduction

Biopharmaceutical products might be recognized by the human immune system as foreign and may induce an unwanted immune response, e.g. the formation of anti-drug antibodies (ADA). Validated assays for the detection and characterization of ADAs are used routinely to monitor a potential onset of immunogenicity. Their formats, testing strategies and performance expectations were described in several regulatory guidelines (Center for Drug Evaluation and Research, 2009; Committee for Medicinal Products for Human Use, 2007; United States Pharmacopeia Convention, 2012) and white papers (Koren et al., 2008; Mire-Sluis et al., 2004; Shankar et al., 2008). Within a multi-tiered approach we focus on the first screening stage, which intends to separate ADA-negative specimens from (true or false) ADA-positive specimens by means of a single screening cut point (SCP). Specifically, we discuss statistical approaches of determining this cut point. Because we consider only the pre-study validation of this screening stage in detail, we use the terms ADA^+ and ADA^- to describe a data-based classification of the in-vitro response in a laboratory assay. At this stage, it is not clear whether a high assay response (classification as ADA^+) for a given method would actually correspond to a clinically relevant ADA reaction. Further, assay results may be prone to artifacts, e.g. single high

* Corresponding author. Tel.: +49 511 762 5821; fax: +49 511 762 4966.
E-mail address: schaarschmidt@biostat.uni-hannover.de (F. Schaarschmidt).

http://dx.doi.org/10.1016/j.jim.2015.02.004
0022-1759/

signals may result from technical errors that still would be considered for cut point calculation. We presume that by using his expertise the responsible bioanalyst reduced technical artifacts to a minimum before starting the pre-study validation exercise. Although not examined in this publication, we suppose that our methods are in principle also useful to re-analyze the correct setting of the cut point when sufficient clinical data become available. To perform such an exercise would be beyond the scope of this publication, and will be in the focus of future research. We only can encourage the interested reader to use the attached R-tool for his own data, in order to see how using a different method for cut point calculation would compare to a currently used cut point.

Method validation of screening assays is typically performed using a set of specimens taken from a drug-naive population (animals or healthy volunteers) of both genders (Shankar et al., 2008). When using a microtiter plate-based approach, usually, duplicated wells are used to obtain one reportable value per specimen, and several negative and positive controls are included on each plate. Furthermore, multiple microtiter plates are needed in order to repetitively analyze the same specimens in several assay runs, and to include other factors, such as between-day heterogeneity, different analysts, or different instruments in the validation exercise. Therefore, normalization of the different plates, e.g. by means of standardization with the plate-specific negative control, is needed. Depending on the assay, this normalization can cause an almost-complete elimination of plate differences, i.e., in such a case a simplified model for data distribution from all runs can be used. However, for the validation exercise of some methods certain differences between plates might remain. This problem can be expected for many assay formats during their development, no matter if microtiter plate-based approaches are used or other methods, e.g. CD- or label-free binding technologies, which also have run and batch-structured data. For some assay methods between run/batch differences cannot be eliminated even after addressing potential underlying reasons by method improvement. In order to deal with such distributions, which might have been caused by a range of pre-existing ADA specificities in the validation population, a more complex random effects model is required for analysis.

Assay methods exist, in which a considerable fraction of ADA^+ specimens is present in the drug-naive sample population (Mikulskis et al., 2013; Tami, 2013; Zhang et al., 2013). Completely ignoring such a phenomenon may lead to screening cut points that are substantially inflated, thus underestimating the number of true positive specimens in the drug-treated study population. Also, simple outlier detection rules like that of the boxplot (i.e., value > 75% percentile + 1.5 × interquartile range, see the outlier-box in Appendix B of Shankar et al., 2008) may fail to exclude all or considerable parts of the ADA^+ specimens if their proportion in the drug-naive sample is large. As an alternative, we propose approaches assuming a mixture distribution of ADA^+ and ADA^- specimens in the drug-naive sample population, as introduced by Jaki et al. (2011).

After the exclusion of potential ADA^+ specimens in the validation population, the next step is the determination of the screening cut point from the set of ADA^- specimens. Based on random effects models, prediction intervals have been proposed as a method for screening cut point estimation (Hoffman and Berger, 2011). Random effects models and corresponding prediction intervals are based on normality assumptions. Thus, an assessment of the adequacy of these assumptions is needed when applying these methods. However, simplistic test procedures like the Shapiro–Wilk test may indicate deviation from normality for several reasons: presence of between plate differences despite normalization, presence of ADA^+ specimens that have not been detected as outliers, or an indeed skewed distribution of the ADA^- specimens that may or may not be remedied by transformations, such as the $\log_e$-transformation. Conversely, when sample sizes are small, test procedures may fail to detect significant deviation from normality. For these reasons, in this article we briefly discuss methods to assess normality after fitting mixture models that may include random effects (Gurka et al., 2007).

Currently, decision-tree approaches are commonly used (Kubiak et al., 2013; Shankar et al., 2008): Depending on the assessment whether outliers are present in the validation data set, and whether normal distribution of the data is given, a certain model for the statistical analysis and calculation of the cut point is selected. However, the automated, unreflecting application of such decision tree approaches may lead to unacceptable screening cut points in case problems as outlined above occur. In the remaining part of the paper, we therefore demonstrate that mixture models, accounting for an unknown, possibly substantial proportion of ADA^+ specimens, can be extended to include random effects, and provide hints at the assessment of the normality assumptions after such model fits. We discuss prediction intervals or non-parametric methods to estimate cut points following the model fits, and illustrate the combined application of all these methods by analyzing two example data sets. However, the application of these methods to particular data sets may be very complex, and oftentimes bioanalysts might not have direct access to statisticians, and are relying on simplified approaches. Indeed, based on empirical experience with multiple assays from different therapeutic areas, simplified cut point estimation can be used in some assays. The challenge is, however, that both scenarios are hard to distinguish in real data. Therefore, we recommend a case-by-case analysis, jointly performed by a bioanalyst and a biostatistician. To illustrate this complex task, the add-on package mixADA for the R software (R Core Team, 2014) is provided.

2. Statistical problems

During development of the described statistical approaches we analyzed roughly 20 data sets obtained during validation of the screening step. These data sets arose from laboratory experiments involving two basic types of factors, i.e. potential sources of variability: 1) a particular set of specimens i.e., sera from several animals or subjects under investigation, 2) different sets of factors representing sources of technical variability, e.g., different analysts or technical devices, different points in time, different plates and repeated observations (usually duplicates) of the same specimen on the same plate. Because these data arose from different assay formats (homogenous or sequential, bridging or sandwich format, ELISA or ECL) developed during different phases in early drug development, the "experimental design" differed with respect to the number of factors and the number of their levels, but also with respect to the nesting or crossing structure among the factors. Although we exemplify such sources of technical variability