Statistical methods for mass spectrometry-based clinical proteomics

Kakourou, A.A.

6

Adapting censored regression methods to adjust for the limit of detection in the calibration of diagnostic rules for clinical mass spectrometry proteomic data

Abstract

We consider the problem of calibrating diagnostic rules based on high-resolution mass-spectrometry (MS) data subject to the limit of detection (LOD). The LOD is related to the limitation of instruments in measuring low-concentration proteins. As a consequence, peak intensities below the LOD are often reported as missings during the quantification step of proteomic analysis. We propose the use of censored data methodology to handle spectral measurements within the presence of LOD, recognizing that those have been left-censored for low-abundance proteins. We replace the set of incomplete spectral measurements with estimates of the expected intensity and use those as input to a prediction model. To correct for lack of information and measurement uncertainty, we combine this

This chapter has been published as: Alexia Kakourou, Werner Vach and Bart Mertens. (2016) Adapting censored regression methods to adjust for the limit of detection in the calibration of diagnostic rules for clinical mass spectrometry proteomic data. *Statistical Methods in Medical Research*, epub ahead of print.

approach with borrowing of information through the addition of an individual-specific random effect formulation. We present different modalities of using the above formulation for prediction purposes and show how it may also allow for variable selection. We evaluate the proposed methods by comparing their predictive performance with the one achieved using the complete information as well as alternative methods to deal with the LOD.

6.1 Introduction

One of the objectives in mass-spectrometry (MS) clinical proteomics is to detect and quantify the proteomic expression in biological samples. Quantification can be affected by measurements being subject to lower detection limits due to censoring mechanisms on low-abundance proteins/peptides (Karpievitch et al., 2009, 2012). This issue is known as limit of detection (LOD) and occurs due to the limited ability of instruments to measure low-concentration proteins. As a result, proteomic data sets, in many applications, consist of reduced lists of peaks with peak intensities below the detection limit threshold reported as missing values.

Several approaches have been proposed in the literature to deal with data subject to (lower or upper) detection limits, particularly in ecological and environmental research (Helsel, 1990, 2012; Hopke et al., 2001; Hornung and D, 2000; Succop et al., 2004). Handling proteomic data subject to LOD has recently emerged in the field of mass-spectrometry clinical proteomics. Dong et al. (2014) addressed the problem of assessing bias in the estimation of distribution parameters of proteomic biomarkers whose measurements were subject to the LOD. They considered a protein pathway data set and proposed methods to combine proteomic markers, adjusting for the LOD, to distinguish cancer from non-cancer patients. They showed that ROC curve parameter estimates generated from the proposed methods are much closer to the truth as compared to simply combining proteomic markers ignoring the LOD. Tekwe et al. (2012) acknowledged the LOD issue in (MS) proteomic data as a problem of censored data analysis and proposed the use of survival methodology, in particular accelerated failure time models (AFT), to investigate differential expression of proteins. They proved that AFT models have higher ability to detect differentially expressed proteins than standard testing procedures, with the discrepancy widening with increasing missingness in the proteomic data.

In this paper, we adapt the use of censored data methodology to handle spectral measurements within the presence of LOD. We implement this approach for the particular problem of estimating the proteomic expression within isotopic clusters with the ultimate goal of using the derived estimates as predictor variables for the calibration of diagnostic rules. In particular, we adapt censored normal regression methods to estimate the expected intensity within an isotope cluster for each individual, based on partially observed MALDI-FTICR mass-spectrometry data, collected in a pancreatic cancer case-control study. The estimates of expected intensity, adjusted for LOD, are later used as input to a prediction model. While censored regression models are widely used in survival analysis

(Therneau and M, 2000; Therneau, 2015; Henningsen, 2010), the specific case of censored normal regression, considered in this paper, is often used in econometrics to handle skewed data and is referred to as Tobit regression (Tobin, 1958). We combine censored regression with borrowing of information, through the addition of an individual-specific random effect formulation, to correct for both lack of information and measurement uncertainty. In addition we present an extension which allows for selection of a subset of features based on the parameter estimates of the censored model. We evaluate the proposed methods by comparing their predictive performance with the one achieved using the complete information as well as alternative methods to deal with the LOD.

The remainder of the paper is organised as follows: In section 6.2 we give a brief overview of the data and the data structure. In section 6.3 we present different frameworks of using random effects censored regression to estimate the average isotope expression in the individual spectra for prediction purposes. Section 6.4 contains results and relative performance of the proposed approaches with ad-hoc methods. Additionally, we show how the use of individual-specific random effects in the censored regression model may allow for the selection of sparse models while maintaining predictive performance. We finish with a discussion in Section 6.5.

6.2 Data

6.2.1 Data description

We consider data from a case-control study, the design of which is described in detail in (Nicolardi et al., 2014). For the experiment, serum samples were collected from 88 patients with pancreatic cancer and 185 healthy volunteers. The samples from the included individuals were stored and processed according to a standardized protocol (Nicolardi et al., 2014). The study design defined a calibration set and a separate validation set. The validation samples were collected in a later time period. For the calibration set, serum samples were obtained from 49 pancreatic cancer patients and 110 healthy controls (age- and gender- matched). For the validation set, samples were obtained from 39 pancreatic cancer patients and 75 healthy (age- and gender- matched) controls. The available calibration and validation samples were distributed over three distinct MALDI-target plates and were mass-analysed by a MALDI-FTICR MS system resulting in a single spectrum per sample covering the mass/charge range from 1013 to 3700 Da.

Figure 6.1 plots the mass spectrum of a single individual (see Kakourou et al. (2016) for another example). A mass spectrum consists of peaks with a certain intensity distributed over a m/z-axis generated from the detected ionized molecules. A single molecule species will appear as a collection of – as opposed to a single – peaks within the proteomic mass spectrum due to the presence of distinct isotopes within the constituent atoms. Isotopes are variants of the same element/atom which occur in nature with different atomic masses due to the presence of additional neutrons in their nucleus. Because of this variation in the numbers of neutrons in the constituent atoms of any given molecule, these

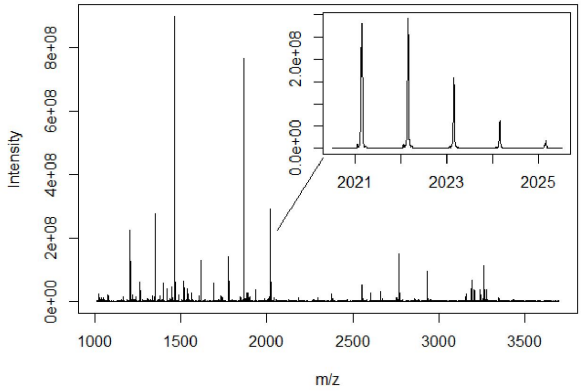

Figure 6.1: The mass spectrum of a single individual. Superimposed is shown (enhanced) an isotopic cluster at position *m/z* 2021,2.

molecules will also occur with varying molecular weight. This in turn causes the spectral proteomic expression for any molecule to appear – not as a single intensity peak at a given mass-to-charge ratio – but rather as a set of neighbouring peaks separated by approximately 1 Dalton, due to the presence or absence of these additional neutrons. We refer to these sets of isotopic peaks as the isotopic cluster . Superimposed in Figure 6.1 is shown an isotopic cluster at position *m/z* 2021,2.

6.2.2 Data structure and limit of detection (LOD)

We apply to the complete raw spectra a peak detection algorithm Kakourou et al. (2016) using a fixed LOD threshold which reflects the background noise level in the individual spectra. This generates a list of 8080 identified isotopic peaks divided into 2717 identified isotopic clusters. In case a peak is observable/detectable in a patient, the approach calculates the area under the intensity curve to obtain an intensity value for that peak and that patient. In case a peak is unobservable/undetectable in a patient, we regard the intensity value for that peak and that patient as left-censored due to the LOD. In Figure 6.2 we plot an identified isotopic cluster at position *m/z* 2021,2 for three different samples to show the possible censoring patterns. The isotopic cluster is completely distinguishable from the background noise in the first sample, only partly distinguishable in the second sample and entirely indistinguishable in the third sample (see Kakourou et al. (2016) for another

Other used terms include isotopic distribution and isotopic envelope.

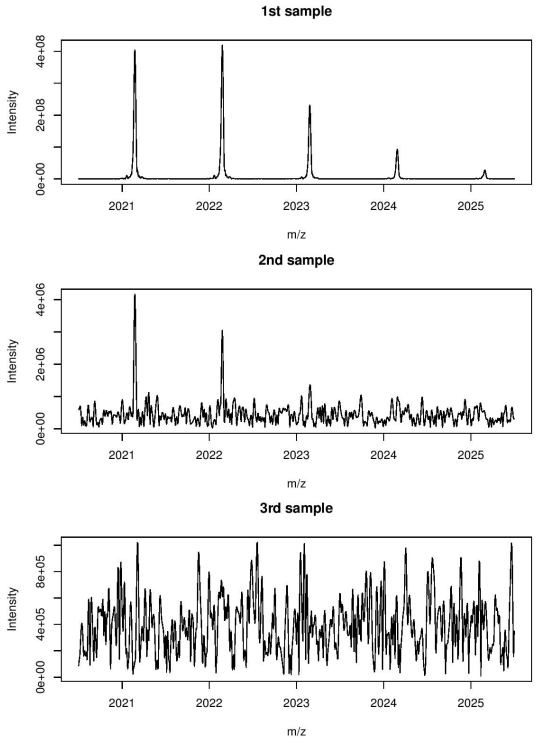

Figure 6.2: An isotopic cluster at position *m/z* 2021,1 for three different samples. The isotopic cluster is completely distinguishable from the background noise in the first sample, only partly distinguishable in the second sample and entirely indistinguishable in the third sample.

example). Note that the overall intensity of the isotopic pattern is decreasing from one sample to the next. Hence, the lower the abundance of a protein/peptide, the less distinguishable the isotopic cluster originating from that molecule is from the background noise. Such differences in abundance could rise, apart from the underlying biological processes, due to variations in the amount of starting material.

The structure of the observed data, given the incomplete response due to the LOD, is given by $(g_i, \mathbf{Y}_i, \mathbf{\Delta}_i)$, $i = 1, ..., n$, where g_i is the group outcome for individual i, $\mathbf{Y}_i$ is a hierarchically structured data matrix consisting of the quantified peak intensity values y_{icj} of peak j in cluster c and $\mathbf{\Delta}_i$ is, analogously, a structured binary matrix representing the censoring indicators which take the value $\delta_{icj} = 1$, if peak j of cluster c is observable, and $\delta_{icj} = 0$, if peak j of cluster c is unobservable in the ith individual. In the latter case, we set the value of y_{icj} to the LOD value.

The output data set contains a large proportion of censored intensity values (85%). Our objective is to investigate whether, starting from the incomplete data, we can develop

methods which will allow us to calibrate diagnostic rules of comparable performance as if we had the complete information.

6.3 Methods

In this section we consider methods to construct, for a given isotopic cluster (the index c is suppressed in the following), estimates $\hat{\bar{y}}_i$ of the average intensity based on the log-transformed peak intensities y_{ij}, with $i = 1, ..., n$ denoting the patients and $j = 1, ..., k$ denoting the peaks of the isotopic cluster. We will use these estimates as input variables for the construction of a diagnostic rule. Our approach is based on a simple model for the intensities in a single patient, which uses the common pattern of the observed intensities across patients as predictor. More specifically, we postulate a regression model for the intensities of a patient in an isotopic cluster, using the empirical pattern of mean intensities $\bar{y}_j := \frac{1}{n}\sum_i y_{ij}$ across patients as covariate. To obtain the estimates of the average intensity for each patient and each isotopic cluster, we use censored regression methodology.

In addition to the censored regression models, we consider some well-known strategies which can deal with unobservable intensity values at the peak level. The first and simplest alternative strategy we consider is complete case analysis, ignoring all censored peak intensities. A simple alternative approach which allows us to use additional information on the unobservable peaks is to reduce all intensity values to the binary information above/below the LOD (Helsel, 2012). Finally, we consider substituting the unobservable peak intensities with the LOD value in order to avoid the loss of information in the observable peaks. For all these alternative methods, we obtain an estimate of the average intensity within an isotopic cluster by averaging the (available) values. For the complete case analysis, if an entire isotopic cluster is unobservable in a sample, we impute the average over the estimates from all patients with at least one observable peak.

6.3.1 Censored regression

We consider for each patient and each isotopic cluster a simple regression model of the type

$$\tilde{y}_{ij} = \alpha_i + \beta_i \bar{y}_j + \varepsilon_{ij} \text{ with } \varepsilon_{ij} \sim N(0, \sigma_i^2)$$

for the true log-transformed intensities $\tilde{y}_{ij}$. The parameters α_i and β_i in the above expression capture the intensity variation of a particular individual. Here, α_i reflects the systematic differences in average expression across patients while β_i represents the (multiplicative) effect of the average isotope pattern $\bar{y}_j$, which is here used as the only predictor of the observed pattern in each individual. The likelihood function based on the partially censored observations is given by

$$L(\theta_i) = \prod_{j=1}^{k} f(y_{ij}, \theta_i)^{\delta_{ij}} F(y_{ij}, \theta_i)^{1-\delta_{ij}}$$

where θ_i is the vector of model parameters, $f(y_{ij}, \theta_i)$ is the probability density function of the normal distribution and $F(y_{ij}, \theta_i)$ is the cumulative distribution function. The contribution of the observed peak intensities to the likelihood is given by $f(y_{ij}, \theta_i)$ while the contribution of the left-censored peak intensities is given by $F(y_{ij}, \theta_i) = Pr(y_{ij} \leq t)$ where t denotes the minimum detectable threshold. The estimates of the regression parameters are obtained by maximizing the log-likelihood function

$$l(\theta_i) = \sum_{j=1}^{k} \delta_{ij} \log f(y_{ij}, \theta_i) + (1 - \delta_{ij}) \log F(y_{ij}, \theta_i)$$

We summarize the entire isotopic expression within each isotopic cluster and for each individual as a function of the estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$, given by

$$\hat{\bar{y}}_i = \frac{1}{k} \sum_{j=1}^{k} \left(\hat{\alpha}_i + \hat{\beta}_i \bar{y}_j \right) = \hat{\alpha}_i + \hat{\beta}_i \bar{\bar{y}}$$

Finally, we use the set of the derived estimates $\{\hat{\bar{y}}_i, \ i = 1, ..., n\}$ as predictors for the construction of the discriminating rule.

6.3.2 Random effect censored regression

Depending on the extent of left-censoring, information in an isotopic cluster for a specific patient may either be insufficient for estimating the model parameters or include great uncertainty resulting in unreliable parameter estimates. We account for lack of information and measurement uncertainty by combining censored regression with shrinkage estimation of the intensity levels. The key idea is to adjust the estimates of the less reliable individual expressions in an isotopic cluster by pooling information across all available patients. In analogy to repeated measures data analysis, we treat the peak intensities within each isotopic cluster for each patient as repeated observations and we fit for each isotopic cluster a joint model across patients including individual-specific random effects.

We restrict our investigation to a simple univariate random effects model, given by

$$\tilde{y}_{ij} = a_i + \alpha + \beta \bar{y}_j + \varepsilon_{ij}$$

with

$$\varepsilon_{ij} \sim N(0, \sigma^2), \ a_i \sim N(0, \tau^2)$$

In the above model specification we choose a fixed effect representation for β, primarily due to computational constraints when fitting a bivariate random effects model, as the fitting process requires numerical integration based on summing up over a number of fixed points which grows exponentially with the number of dimensions. In this way, α represents the mean intercept across all patients while a_i represents the individual deviation from the mean. The variation of the individual intercepts around the mean is assumed

to be normally distributed.

The likelihood function for the parameter vector $\theta = \left(\alpha, \beta, \tau^2, \sigma^2\right)$ is given by

$$L(\theta; \mathbf{y}_i) = \prod_{i=1}^{n} \left(\int_{-\infty}^{+\infty} \prod_{j=1}^{k} \left(f_\theta(y_{ij}|a_i)^{\delta_{ij}} F_\theta(y_{ij}|a_i)^{1-\delta_{ij}} \right) f_\theta(a_i)\, da_i \right)$$

where $\mathbf{y}_i$ denotes the intensity vector which may include one or more undetectable values. When undetectable peak intensities are to be accounted for, the Bayes estimate of the random effect a_i can be computed by substituting the ML estimates of $\theta = \left(\alpha, \beta, \tau^2\sigma^2\right)$ into the analytic expression for the posterior mean given the observed data, given by

$$E(a_i|\mathbf{y}_i, \theta) = \frac{1}{f^*(\mathbf{y}_i; \theta)} \int_{-\infty}^{+\infty} a_i\, f^*(\mathbf{y}_i|a_i) f(a_i)\, da_i$$

where,

$$f^*(\mathbf{y}_i, \theta) = \int_{-\infty}^{+\infty} f^*(\mathbf{y}_i, a_i) da_i = \int_{-\infty}^{+\infty} f^*(\mathbf{y}_i|a_i) f(a_i)\, da_i$$

and

$$f^*(\mathbf{y}_i|a_i) = \prod_{j=1}^{k} \left(f_\theta(y_{ij}|a_i)^{\delta_{ij}} F_\theta(y_{ij}|a_i)^{1-\delta_{ij}} \right)$$

The above expression for the random intercept estimate is equivalent to

$$\hat{a}_i = E(a_i|\mathbf{y}_i, \theta) = \frac{\tau^2}{\tau^2 + \sigma^2/k} \left(E\left(\bar{\tilde{y}}_i|\mathbf{y}_i, \boldsymbol{\delta}_i, \hat{\theta}\right) - \left(\hat{\alpha} + \hat{\beta}\bar{\tilde{y}}\right) \right)$$

as pointed out by Hughes (1999), who proposed prediction of random effects in conjunction with an EM approach to the LOD problem (see Appendix A for detailed derivation). The Bayes estimate $\hat{a}_i$ can be considered as a weighted average of the prior mean of a_i which is assumed 0 and the average residual $\left(E\left(\bar{\tilde{y}}_i|\mathbf{y}_i, \boldsymbol{\delta}_i, \hat{\theta}\right) - \left(\hat{\alpha} + \hat{\beta}\bar{\tilde{y}}\right) \right)$. Finally, the estimated intensities for the ith individual can be written as

$$\begin{aligned} \hat{y}_{ij} &= \hat{\alpha} + \hat{\beta}\bar{y}_j + \hat{a}_i \\ &= \hat{\alpha} + \hat{\beta}\bar{y}_j + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2/k} \left(E\left(\bar{\tilde{y}}_i|\mathbf{y}_i, \boldsymbol{\delta}_i, \hat{\theta}\right) - \left(\hat{\alpha} + \hat{\beta}\bar{\tilde{y}}\right) \right) \end{aligned}$$

Viewed this way, the estimate of τ^2 plays the role of a shrinkage parameter. The smaller the value of τ^2 relative to σ^2, the more the shrinkage of the random effect estimate of a particular individual towards zero and hence the more the shrinkage of the intensity estimates of that individual towards the population mean. Note that $\hat{y}_{ij}$ is defined also in the case of completely unobservable isotopic clusters, hence patients for which all intensity values are censored in an isotopic cluster can also be included in the analysis.

6.3.3 Random effect censored regression applications

There are several possibilities of using the random effect censored regression model to estimate the individual isotopic expression with the ultimate goal of using the derived estimates as predictors in the calibration of the diagnostic rule. In the following we demonstrate three different variants of using the random effect censored regression model, for prediction purposes.

Random effect censored regression as a preprocessing tool

A simple and straightforward way to summarize the incomplete predictive information in the pancreatic data, while accounting for the LOD, is to apply the random effect censored regression approach across all available data i.e. data from both calibration and validation sets. This can be considered as a means of preprocessing the data, without using information on the class outcome, prior to building the diagnostic rule.

The Bayes estimate of the random intercept in this case is given by $\hat{a}_i(\hat{\theta}_{all}) = E(a_i|\mathbf{y}_i, \theta_{all})$ where $\hat{\theta}_{all} = (\hat{\alpha}_{all}, \hat{\beta}_{all}, \hat{\tau}^2_{all}, \hat{\sigma}^2_{all})$ and *all* denotes the fact that, conditioning on the random effect, the the estimates of the parameter vector were derived based on all the available observations. Correspondingly, the expected intensity for patient i and peak j is given as

$$\hat{y}_{ij}(\hat{\theta}_{all}) = \hat{a}_i(\hat{\theta}_{all}) + \hat{\alpha}_{all} + \hat{\beta}_{all}\, \bar{y}_j$$

while the average expected intensity within isotopic cluster for patient i is derived by

$$\hat{\bar{y}}_i(\hat{\theta}_{all}) = \hat{a}_i(\hat{\theta}_{all}) + \hat{\alpha}_{all} + \hat{\beta}_{all}\, \bar{\bar{y}}$$

Random effect censored regression as a prediction tool

A more formal approach, more in tune with predictive calibration and subsequent validation, is to embed the above estimation procedure within the ordinary prediction framework. This suggests using the calibration data for both parameter estimation of the random censored model and construction of the prediction model and subsequently applying the resulting rules to the set-aside validation set.

In that case, the Bayes estimate of the random intercept is given by $\hat{a}_i(\hat{\theta}_{cal}) = E(a_i|\mathbf{y}_i, \theta_{cal})$ with $\hat{\theta}_{cal} = (\hat{\alpha}_{cal}, \hat{\beta}_{cal}, \hat{\tau}^2_{cal}, \hat{\sigma}^2_{cal})$ where *cal* denotes the fact that the parameter estimates were based solely on the calibration samples. In other words, both calibration and validation data are shrunken according to the estimates derived based on the calibration set alone. The expected intensity of each patient i, for the calibration and the validation sets, is given by

$$\hat{y}_{ij_{cal}}(\hat{\theta}_{cal}) = \hat{a}_{i_{cal}}(\hat{\theta}_{cal}) + \hat{\alpha}_{cal} + \hat{\beta}_{cal}\, \bar{y}_j \quad \text{and} \quad \hat{y}_{ij_{val}}(\hat{\theta}_{cal}) = \hat{a}_{i_{val}}(\hat{\theta}_{cal}) + \hat{\alpha}_{cal} + \hat{\beta}_{cal}\, \bar{y}_j$$

respectively while the corresponding isotopic cluster summary for each set is given by

$$\hat{\bar{y}}_{i_{cal}}(\hat{\theta}_{cal}) = \hat{a}_{i_{cal}}(\hat{\theta}_{cal}) + \hat{\alpha}_{cal} + \hat{\beta}_{cal}\, \bar{\bar{y}} \quad \text{and} \quad \hat{\bar{y}}_{i_{val}}(\hat{\theta}_{cal}) = \hat{a}_{i_{val}}(\hat{\theta}_{cal}) + \hat{\alpha}_{cal} + \hat{\beta}_{cal}\, \bar{\bar{y}}$$

Random effect censored regression re-estimation

Estimating the average intensity in the validation data, either in conjunction with the calibration data, as in the case of the random censored regression model as a preprocessing tool, or according to the censored regression estimates based on the calibration data, as in the case of the random censored regression model as a prediction tool, implies that one assumes that both calibration and validation data stem from the same population.

If the above assumption does not hold it may be necessary and/or beneficial to estimate the censored model parameters separately for the calibration and validation data since difference in population could result in potentially different values of α, β, τ^2 or σ^2 between the two sets. This could be particularly true in the case of external validation where validation samples may represent a different population than calibration samples, for instance due to population drift (Kelly et al., 1999; Adams et al., 2010).

On the assumption that the two populations are different, we fit the random effect censored regression model separately to the calibration and the validation data. In this case, the random effect estimate for the calibration samples is given by $\hat{a}_{i_{cal}}(\hat{\theta}_{cal}) = E(a_{i_{cal}}|y_{i_{cal}},,\theta_{cal})$, where $\theta_{cal} = \left(\alpha_{cal}, \beta_{cal}, \tau^2_{cal}, \sigma^2_{cal}\right)$, while the random effect estimate for the validation samples is given by $\hat{a}_{i_{val}}(\hat{\theta}_{val}) = E(a_{i_{val}}|y_{i_{val}},,\theta_{val})$, where $\theta_{val} = \left(\alpha_{val}, \beta_{val}, \tau^2_{val}, \sigma^2_{val}\right)$. The resulting intensity estimates for the calibration and validation sets are derived by

$$\hat{y}_{ij_{cal}}(\hat{\theta}_{cal}) = \hat{a}_{i_{cal}}(\hat{\theta}_{cal}) + \hat{\alpha}_{cal} + \hat{\beta}_{cal}\,\bar{y}_j \quad \text{and} \quad \hat{y}_{ij_{val}}(\hat{\theta}_{val}) = \hat{a}_{i_{val}}(\hat{\theta}_{val}) + \hat{\alpha}_{val} + \hat{\beta}_{val}\,\bar{y}_j$$

respectively while their corresponding isotopic cluster summaries are defined as

$$\hat{\bar{y}}_{i_{cal}}(\hat{\theta}_{cal}) = \hat{a}_{i_{cal}}(\hat{\theta}_{cal}) + \hat{\alpha}_{cal} + \hat{\beta}_{cal}\,\bar{\bar{y}} \quad \text{and} \quad \hat{\bar{y}}_{i_{val}}(\hat{\theta}_{val}) = \hat{a}_{i_{val}}(\hat{\theta}_{val}) + \hat{\alpha}_{val} + \hat{\beta}_{val}\,\bar{\bar{y}}$$

6.4 Application and analysis

6.4.1 Model choice

We assess the performance of the proposed methods by fitting a prediction model to the set of the derived isotopic cluster summaries and by evaluating the predictive performance of each fit. Summarizing the isotopic expression per cluster results in a total of 2717 isotopic cluster summaries, reducing the dimensionality of the original predictor data. As the number of predictors still exceeds the number of observations, we choose ridge logistic regression (Le Cessie and van Houwelingen, 1992) to calibrate the diagnostic rule. This method is effective in high-dimensional settings, where the number of covariates exceeds the number of observations and/or there are high correlations between them. Ridge regression deals with overfitting and collinearity by maximizing the log-likelihood function with a penalty on the regression coefficients.

6.4.2 Model fitting and performance measures

To fit the random intercept censored regression model we use the NLMIXED procedure available in SAS which is written to fit non-linear mixed models (SAS code is given in Appendix B1). The computation of the integral over the random effect is performed by an adaptive Gaussian quadrature method with 100 integration points. Since in our data analysis we consider integrated intensities, the original LOD value used in our peak detection algorithm is no longer adequate, as it only indicates that the maximal intensity in an interval around the peak is below that value. Taking into account that the width and shape of a particular peak is approximately constant across patients, we decided to use as peak-specific LOD value, the minimal observed integrated intensity among all patients with an uncensored measurement.

To evaluate the proposed methods with respect to their predictive performance, we first apply an internal validation in which we use random splitting to redefine the calibration and validation sets. This allows us to assess consistency of performance estimates and obtain more robust results. The new calibration-validation structure is defined in such a way that it respects the case/control ratio of the original study design (see Kakourou et al. (2016) for detailed description). The procedure is repeated 10 times and classification results across the repetitions are averaged to obtain more stable estimates.

To choose the optimal value of the ridge penalty we perform leave-one-out cross validation on the re-defined calibration set. The resulting classification rule is then evaluated on the re-defined validation set. For each model we calculate the error-rate and the area under the ROC curve (AUC). To evaluate the accuracy of each fit, we calculate the Brier score, given by

$$Brier\ score \quad = \quad \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} (\hat{p}_i - g_i)^2$$

and the deviance, defined as

$$\begin{aligned} Deviance \quad &= \quad -2 \sum_{i=1}^{n_{val}} g_i \log \hat{p}_i + (1 - g_i)(\log (1 - \hat{p}_i) \\ &= \quad -2 \sum_{i=1}^{n_{val}} \log(1 - |\hat{p}_i - g_i|) \end{aligned}$$

where $\hat{p}_i$ is the estimated probability of being a case for the i^{th} validated individual, g_i is the known class outcome of that individual and n_{val} is the total validation sample size. To compute the error-rates we use a threshold of 0.5 and we assign an observation as a diseased case if the predicted class probability $\hat{p}_i$ is greater than 0.5 and as a control otherwise.

6.4.3 Results

Table 6.1 shows validated performance measures, together with standard errors, of the ridge logistic model fitted to the set of average estimates based on complete case analysis (CCA), binary coding (BC), substituting unobservable peak intensities with the detection limit value (LOD), random censored regression as a preprocessing tool (CR Prep), as a prediction tool (CR Pred) and re-estimated (CR Reest). The last column of Table 6.1 contains performance measures based on substituting the unobservable peak intensities with the area under the intensity curve in a symmetric interval around the peak position, with length corresponding to the typical peak width in a specific *m/z* range, estimated from the raw data. This approach can be considered equivalent to having the complete information on the peak intensities (the "truth") and it is feasible in our specific situation since we have access to the complete spectra and not just a peak list as is often the case. Therefore, assessment of relative performance may be carried out with respect to this approach (TR).

Performance measures, based on BC suggest that the present/absent patterns of the proteomic expression are highly informative with regards to the class outcome. Incorporating additional information on the relative intensity, while accounting for the LOD, seems to be recovering information on top of the present/absent information. Specifically, results based on CR Prep or CR Pred indicate that using censored regression strategies combined with pooling of information can solve the LOD problem, both from a statistical and practical point of view. Performance of CR Reest shows, as expected, no improvement over CR Prep or CR Pred, as, in the case of internal validation, calibration and

	Validated classification results (based on internal validation)						
	CCA	BC	LOD	CR Prep	CR Pred	CR Reest	TR
Error-rate	0.135	0.125	0.114	0.110	0.109	0.114	0.114
	(0.008)	(0.007)	(0.005)	(0.005)	(0.006)	(0.008)	(0.005)
Brier score	0.103	0.100	0.085	0.086	0.084	0.086	0.087
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.003)
Deviance	55.70	54.59	46.56	47.41	47.23	48.12	48.08
	(1.85)	(2.18)	(2.00)	(1.89)	(2.01)	(2.16)	(1.67)
AUC	0.917	0.917	0.943	0.940	0.942	0.942	0.940
	(0.006)	(0.009)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)

Table 6.1: Validated classification results (and standard errors) based on complete case analysis (CCA), binary coding (BC), LOD imputation (LOD), random censored regression as preprocessing tool (CR Prep), random censored regression as prediction tool (CR Pred), random censored regression re-estimated (CR Reest) and the "truth" (TR).

validation populations cannot differ systematically.

Results based on CCA illustrate that ignoring the unobservable peak intensities results in poor classification results as compared to results based on methods designed for censored data. Interestingly, we observe that substituting the unobservable peak intensities with the LOD value results in comparable performance to the one achieved using CR Prep, CR Pred,CR Reest or TR. However, this is not utterly surprising as the LOD value for the pancreatic cancer data is a rather accurate estimate of the true (unobservable) intensity value.

Next, we apply each method to the original data.We regard this as our external validation where we use the estimates from the calibration set, as defined in the original study, to build the classification rule and the estimates from the validation set to assess the predictive performance of the derived rule. Validated classification results for all methods are shown in Table 6.2. With one exception, we observe comparable ranking of the methods with the one based on internal validation. Improvement in predictive performance of the proposed censored regression methods as compared to CCA (as well as all other methods including TR) is more apparent in this case, as indicated by both the error-rate and the AUC. In particular, CR Reest now outperforms all methods (including CR Prep, CR Pred and TR) in all performance measures. This outcome provides some confirmation on the value of the re-estimation approach when the two populations are known to be different, as in the case of our external validation. Investigations which would allow us to gain more insight into the possible situations under which the re-estimation approach is expected to outperform the alternative strategies is left as an interesting line of future research.

Finally, we explore to which degree the achieved classification performance when using random censored regression as a solution to the LOD problem is due to borrowing of information or due to shrinkage of the level estimates. We address this question by fitting a prediction model with $E\left(\bar{\tilde{y}}_i|\mathbf{y}_i, \boldsymbol{\delta}_i, \hat{\theta}\right)$ as input variable. In case of no censoring, the above expression reduces to the observed average intensity within the isotopic cluster. In this way we allow for "borrowing" in estimating the average intensity of an isotopic

	Validated classification results (based on external validation)						
	CCA	BC	LOD	CR Prep	CR Pred	CR Reest	TR
Error-rate	0.135	0.125	0.115	0.087	0.096	0.076	0.106
Brier score	0.113	0.107	0.082	0.082	0.082	0.064	0.079
Deviance	78.01	70.50	58.03	56.99	57.64	47.24	54.62
AUC	0.905	0.939	0.956	0.970	0.967	0.971	0.970

Table 6.2: Validated classification results based on complete case analysis (CCA), binary coding (BC), LOD imputation (LOD), random censored regression as preprocessing tool (CR Prep), random censored regression as prediction tool (CR Pred), random censored regression re-estimated (CR Reest) and the "truth" (TR).

	Validated classification results					
	Internal validation			External validation		
	CR Prep	CR Pred	CR Reest	CR Prep	CR Pred	CR Reest
Error-rate	0.115 (0.007)	0.115 (0.005)	0.119 (0.010)	0.115	0.086	0.086
Brier score	0.089 (0.004)	0.085 (0.003)	0.094 (0.007)	0.085	0.084	0.081
Deviance	49.89 (2.16)	48.23 (2.24)	50.19 (4.09)	60.12	59.34	56.95
AUC	0.934 (0.005)	0.939 (0.006)	0.939 (0.006)	0.964	0.959	0.967

Table 6.3: Validated classification results based on random censored regression with no shrinkage as preprocessing tool (CR Prep), random censored regression with no shrinkage as prediction tool (CR Pred), random censored regression with no shrinkage re-estimated (CR Reest) for internal validation (left part) and external validation (right part).

cluster without using shrinkage. Performance measures using the "unshrunken" estimates derived based on CR Prep, CR Pred and CR Reest for the internal and external validations are shown in Table 6.3. Looking at Tables 6.1 and 6.2, we observe that, in the case of internal validation, using censored regression methods with and without shrinkage are of comparable performance. This suggests that the use of censored regression can solve the LOD problem also when it is not combined with shrinkage. In the case of external validation, we observe larger discrepancies favouring the use of shrinkage. This outcome suggests that there may be situations where using shrinkage has a value in its own right.

6.4.4 Variable selection

Combining censored regression with borrowing of information may allow for some type of variable selection, based on the estimate of the random effect variance. As already discussed in Section 6.3.2, the variance of the cluster-specific random intercept τ_c^2 acts as a shrinkage parameter. Depending on the amount and reliability of the available information in an isotopic cluster, the estimates of a specific individual are pulled to a smaller or a greater extent, towards the common population mean. Accordingly, the larger the value of τ_c^2, the higher the spread from patient to patient and hence the more informative that cluster may be. The above consideration can be used as a criterion to eliminate isotopic clusters with minimal τ_c^2. Already, the random effect variances for 15 isotopic clusters were estimated as 0 by the CR Prep approach and thus these clusters were automatically ignored by ridge regression.

Selecting a subset of isotopic clusters, while maintaining predictive performance, can be of particular interest for potentially measuring solely proteins at predefined *m/z* locations. This could reduce the cost of measurement and storage for future data and facilitate

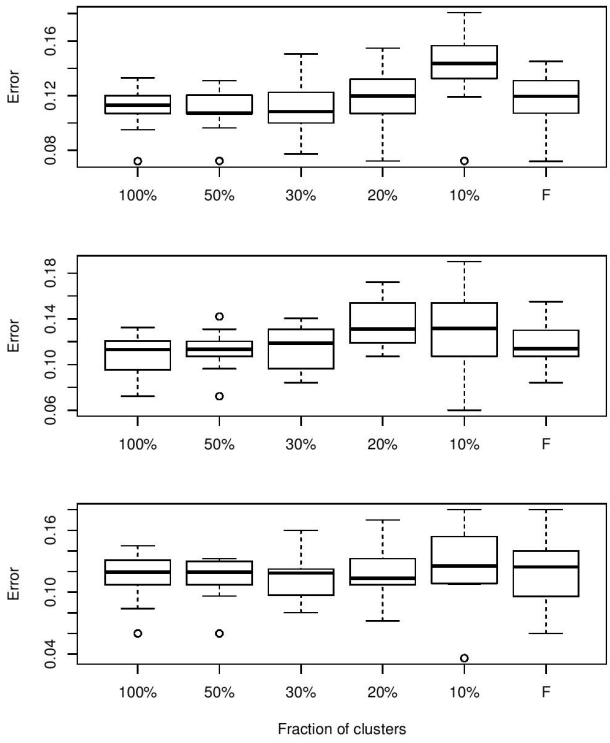

Figure 6.3: Boxplots of error-rates for the 10 re-sampled validation sets when keeping all clusters, 50%, 30%, 20%, 10% and F (=optimal fraction of selected clusters defined based on cross-validation) of total clusters with minimal τ_c^2 based on CR Prep (upper plot), CR Pred (middle plot) and CR Reest (lower plot).

all subsequent analyses. Moreover, variable selection may allow for the identification of a set of features associated with the disease mechanisms and therefore could provide leads to further exploit diagnostic and therapeutic potential.

Variable selection prior to calibration

A simple way to perform variable selection is to discard a certain fraction of isotopic clusters with minimal τ_c^2. For instance, we may decide to eliminate 50%, 80% or 90% of the total number of isotopic clusters with minimal τ_c^2 prior to calibrating the diagnostic rule. In this way, the decision on which isotopic clusters to retain or omit depends solely on the magnitude of the random effect variance and not on cross-validated risk. The first 5 boxplots of Figure 6.3 represent validated error-rate distributions for the 10 re-sampled internal validation sets when keeping all, 50%, 30%, 20% and 10% of total isotopic clusters with minimal τ_c^2, as estimated by CR Prep (upper plot), CR Pred (middle

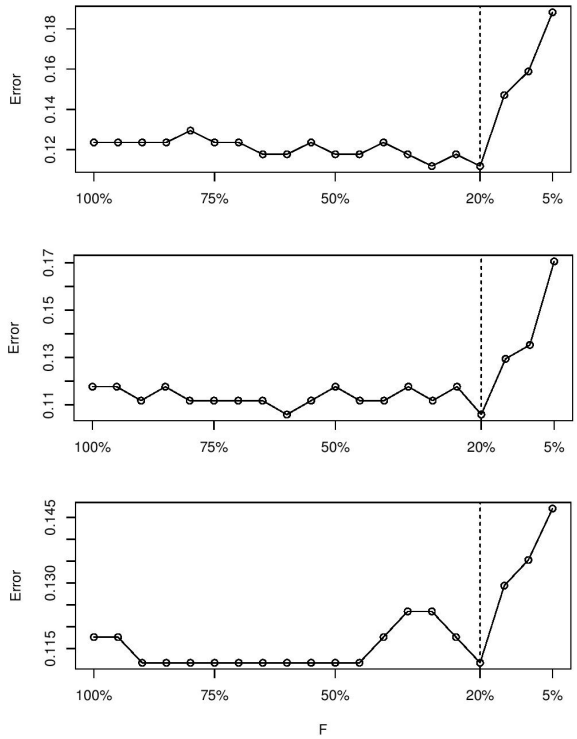

Figure 6.4: Cross-validated error-rates on a re-sampled calibration set for CR Prep (upper plot), CR Pred (middle plot) and CR Reest (lower plot) as the fraction of selected clusters F becomes smaller. Optimal solution is chosen for $F = 20\%$ resulting in a subset of just 543 clusters/proteins.

plot) and CR Reest (lower plot). These results suggest that we can omit at least half of the isotopic clusters from the analysis without deteriorating the predictive performance. Note that for CR Reest, though we get different regression estimates for the calibration and validation sets, the decision on which isotopic clusters to omit is based solely on the random effect variance estimate of the calibration set.

Variable selection within calibration

Alternatively, we may choose the optimal fraction of selected isotopic clusters directly from the predictive perspective via cross-validation. We do so by considering the fraction of selected isotopic clusters F as a tuning parameter to be optimized. In this case, estimation involves combined optimization of the fraction F and the ridge penalty λ. To optimize F (in conjunction with λ) we perform leave-one-out cross-validation on the calibration set for a grid of 20 F values corresponding to the ventiles of $\boldsymbol{\tau}^2 = (\tau_1^2, ..., \tau_C^2)$

(R code is given in Appendix B2).The final classification rule for optimal F and λ is evaluated on the validation set.

Figure 6.4 shows cross-validated error rates for a single re-sampled internal calibration set as the fraction of selected isotopic clusters F becomes smaller for CR Prep (upper plot), CR Pred (middle plot) and CR Reest (lower plot). The cross-validated error, in all cases, is minimized at $F = 20\%$, resulting in a subset of just 543 isotopic clusters (proteins). If the curve is flat near the minimum, we choose the smallest fraction that achieves the minimal error, favoring hence sparser models. The validated error-rate distributions based on optimal F for the 10 re-sampled internal validation sets and three variants are shown in the last boxplots of Figure 6.3. The cross-validated estimates of F which were typically selected across the internal validation sets and investigated methods were either 30% or 20%.

6.5 Discussion

In this paper we proposed to adapt censored regression methods to estimate the average individual expression within isotopic clusters, prior to building prediction rules, as a way to deal with the limit of detection. We evaluated the proposed methods, with respect to predictive performance, by replacing the incomplete spectral measurements with the derived estimates of individual expression, accounted for the LOD, and using those as predictors for the construction of diagnostic rules. We combined censored regression with borrowing of information across data to account for potential lack of information and measurement uncertainty. Results from both internal and external validations indicated that using the estimates from the proposed methods as input variables results in comparable predictive accuracy to the one achieved using the complete intensity information. Ignoring the unobservable peak intensities, as an alternative to deal with the LOD, resulted in poor predictions as compared to the proposed methods, while substituting the unobservable peak intensities with the LOD value exhibited similar classification performance as the proposed methods.

We demonstrated different variants of using censored regression, in combination with borrowing of information, for prediction purposes. Random censored regression as a preprocessing tool is straightforward in application since it only requires fitting the random censored model across all available data. However, since the derived estimates of a particular individual depend now on the expression from all other individuals due to the explicit borrowing of information, information from the validation samples enters the rule derived based on the calibration samples. Since our objective is prediction, we may choose to avoid this by using only the calibration samples to fit the censored regression model and use the derived estimates to adjust for the LOD in the validation set. This approach respects the formal prediction framework.

Another aspect related to the above comparative discussion between censored regression as a preprocessing or prediction tool is the potential need of re-estimating the random censored model parameters in the validation set when the samples represent a different

population than the calibration samples, as in the case of external validation. In classification problems the issue of population difference may not be as crucial, since we model the conditional distribution of the class outcome. However, in data preprocessing, where we model each single univariate covariate, population difference could result in distinct parameter estimates when applying the random effect censored regression model separately on the calibration and validation data. Results based on our external validation provided evidence that the re-estimation approach is superior in this case. Nevertheless, further investigations are required to find out in which situations and to which degree one can benefit from choosing this approach to account for the LOD.

We restricted our discussion to the simple case of the univariate random effects model with random intercept only. We chose to use the univariate random effects model due to its relative ease in computation as opposed to the bivariate case with both random intercept and slope, as fitting these models requires numerical integration based on summing up over a number of fixed grid points which grows exponentially with the number of dimensions. In fact, it might be of interest to consider the bivariate random effects model since the degree to which the average pattern is predictive of the observed pattern may vary from patient to patient. However, results based on using the estimates from a bivariate random censored model (as a preprocessing tool) as predictors were identical to those based on using the estimates from the univariate random censored model (as a preprocessing tool), suggesting that incorporating this additional information does not improve predictions. Moreover, for a large number of isotopic clusters, the estimate of the random slope variance was close to zero, justifying thus the choice of keeping the slope fixed.

A property of using random effect censored regression methods to estimate the expected isotopic expression is that it offers the possibility to use the estimate of the random effect variance as a criterion to screen out the most interesting proteins associated with the disease mechanisms. We presented two different approaches for performing variable selection based on the estimate of the random effect variance. In the first approach, the decision on which variables to retain is based solely on the magnitude of the random intercept variance τ^2 and is independent of the class outcome. On the other hand, one can choose to determine the optimal fraction of variables to be retained by optimizing the loss function through the use of cross-validation, as demonstrated in section 6.4.4.2. Variable selection methods based on optimizing a chosen risk function by looking at the class outcome have seen many applications and publications, with Lasso regularization being among the most popular ones. A formal comparison between the various variable selection methods and the here proposed approach falls beyond the scope of this work.

Apart from *a priori* variable selection based either on a fixed fraction of isotopic clusters with minimal τ^2 or on selecting the optimal fraction of clusters to be retained through the use of cross validation, one could think of alternative ways to determine a reasonable, fixed across isotopic clusters, value for τ^2 in order to get closer to the idea of prediction. One option towards that direction would be to treat τ^2 as a tuning parameter to be optimized via cross-validation. More specifically, rather than estimate τ^2 through maximum likelihood estimation we could consider τ^2 as a fixed parameter in the censored regression

model and perform a grid search to optimize this parameter with respect to predictive performance. In this way, the amount of shrinkage of the intensity levels is estimated directly from a predictive point of view. We leave the idea of determining the optimal value of the random effect variance via optimization techniques, as an interesting topic of future research.

6.6 Conclusion

We have demonstrated that censored regression can be used successfully to handle the LOD problem in determining the average intensity of isotopic clusters in mass-spectrometry proteomic data. In particular in combination with random effects methodology it can contribute to a more efficient preprocessing.

Appendix

A. Derivation of expression for the estimate of the random intercept in the censored regression model

We consider the standard mixed effects model (Laird and Ware, 1982)

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i b_i + \epsilon_i \tag{1}$$

where $\mathbf{Y}_i$ is the response vector (in our case this corresponds to the vector of intensity estimates in an isotopic cluster) for individual i, $i = 1, .., n$, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are design matrices, β is the vector of fixed effects, b_i is the vector of random effects for individual i and ϵ_i is the vector containing the residual components. We assume that b_i and ϵ_i are independent with

$$b_i \sim N(0, T)$$
$$\epsilon_i \sim N(0, \sigma^2\mathbf{I})$$

The Bayes estimate for b_i is then given by

$$\hat{b}_i = E[b_i|Y_i = y_i] = \int b_i f(b_i|y_i)db_i = \mathbf{T}\mathbf{Z}_i'\mathbf{W}_i(y_i - \mathbf{X}_i\beta) \tag{2}$$

where $\mathbf{W}_i = \mathbf{V}_i^{-1}$, $\mathbf{V}_i = var(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{T}\mathbf{Z}_i' + \sigma^2\mathbf{I}$ and $\mathbf{W}$ has a block diagonal structure.

Hughes (1982) showed that the approach by Laird and Ware for estimating b_i can be extended to the case where $\mathbf{Y}_i$ is incompletely observed due to left or right censoring imposed by lower or upper detection limits. In that case, he showed that estimates of b_i may be computed as

$$\hat{b}_i = \mathbf{T}\mathbf{Z}_i'\mathbf{W}_i(E(\widetilde{\mathbf{Y}}_i, |\mathbf{Y}_i, \delta_i, \theta) - \mathbf{X}_i\beta) \tag{3}$$

where δ_i is the censoring indicator.

Now consider the special case of a random-intercept model. The random-effects covariance matrix $\mathbf{T}$ reduces to a scalar corresponding to the variance of the random intercept which we denote by τ^2. The design matrix $\mathbf{Z}_i$ is now a k-dimensional vector of ones where k is the number of responses/peak intensities for the ith individual. We denote that by $\mathbf{1}_k$ and we denote $\mathbf{1}_k\mathbf{1}_k'$ by $\mathbf{J}_k$ (Verbeke and Molenberghs, 2000). It follows from (3) that the estimate for the random intercept of individual i, when $\mathbf{Y}_i$ is incompletely

observed, is given by

$$\begin{aligned}\hat{b}_i &= \tau^2 \mathbf{1}_k'(\tau^2 \mathbf{J}_k - \sigma^2 \mathbf{I}_k)^{-1}(E(\widetilde{\mathbf{Y}}_i, |\mathbf{Y}_i, \delta_i, \theta) - \mathbf{X}_i\beta) \\ &= \frac{\tau^2}{\sigma^2}\mathbf{1}_k'(\mathbf{I}_k - \frac{\tau^2}{\sigma^2+\tau^2}\mathbf{J}_k)(E(\widetilde{\mathbf{Y}}_i, |\mathbf{Y}_i, \delta_i, \theta) - \mathbf{X}_i\beta) \\ &= \frac{\tau^2}{\tau^2+\sigma^2/k}\frac{1}{k}\sum_{j=1}^{k}(E(\tilde{y}_{ij}|y_{ij}, \delta_{ij}, \theta) - x_{ij}'\beta) \\ &= \frac{\tau^2}{\tau^2+\sigma^2/k}(E(\bar{\tilde{y}}_i|y_i, \delta_i, \theta) - \bar{x}_i\beta)\end{aligned} \tag{4}$$

Since we consider as our only predictor the empirical pattern of mean intensities $\bar{y}_j := \frac{1}{n}\sum_{i=1}^{n} y_i$ across patients, the above expression takes the form

$$\hat{b}_i = \frac{\tau^2}{\tau^2+\sigma^2/k}(E(\bar{\tilde{y}}_i|y_i, \delta_i, \theta) - (\alpha + \beta\bar{\bar{y}})) \tag{5}$$

where α denotes the fixed-effects intercept and β denotes the fixed-effects slope.

B. Software implementation

B1. SAS code to fit the random-intercept censored regression model

For fitting the random-intercept censored regression models presented in this article, SAS code using PROC NLMIXED is provided for the three different variants (CR Prep, CR Pred and CR Re-est). Note that in order to use PROC NLMIXED, the data for the analysis must be in long format.

We remind that, conditioning on the random intercept, the contribution of an observed peak intensity to the likelihood is given by the normal probability density function

$$f(y_{ij}|a_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{\frac{-(y_{ij}-\mu_{ij})^2}{2\sigma^2}}$$

while the contribution of a left censored peak intensity is the cumulative density function

$$F(y_{ij}|a_i) = P(y_{ij} \leq t|a_i) = \phi\left(\frac{y_{ij}-\mu_{ij}}{\sigma}\right)$$

where

$$\mu_{ij} = E(y_{ij}|a_i) = a_i + \alpha + \beta\bar{y}_j$$

and t denotes the minimum detectable threshold.

The likelihood function for partially observed data is then given by

$$\mathcal{L}(\theta) = \begin{cases} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{\frac{-(y_{ij}-\mu_{ij})^2}{2\sigma^2}} & \text{if } y_{ij} > t \\ \phi\left(\frac{y_{ij}-\mu_{ij}}{\sigma}\right) & \text{if } y_{ij} \leq t \end{cases}$$

We use the above formulation to specify the likelihood function with PROC NLMIXED and to model the censored intensities using both fixed and random effects. The following lines of code present how to estimate y_{ij} based on the CR Prep approach using PROC NLMIXED.

```
/* Read FTMS_long data into SAS, and create a new
   dataset (ftms_long) in long format */

data ftms_long;
   infile 'FTMS_long.txt';
   input cluster id y lod ybar;
run;

/* Fit the random-intercept censored regression model
   across all samples and for each cluster separately */

proc nlmixed data=ftms_long XTOL=1E-12 method=GAUSS qpoints=100;
   parms alpha=0.7 beta=5 sigma2=0.7 tau2=0.5;
   bounds sigma2 tau2 >= 0;
   pi = constant('pi');
   mu = alpha + a_i + beta*ybar;
   if y > lod then
   ll = (1/(sqrt(2*pi*sigma2)))*exp(-(y-mu)**2/(2*sigma2));
   if y <= lod then
   ll = probnorm((y-mu)/sqrt(sigma2));
   L=log(ll);
   model y ~ general(L);
   random a_i ~ normal(0, tau2) subject = id;
   predict mu out=yexp_all(KEEP = pred);
   by cluster;
run;
```

Next we present how to fit the random-intercept censored regression model based on the CR Pred approach. To estimate $\hat{\hat{y}}_{ij(cal)}$ the code is the same as before except we only use the data from the calibration set and we save the estimates of α (`alpha`), β (`beta`), σ^2 (`sigma2`) and τ^2 (`tau2`) so that they can be used later as fixed variables in order to estimate $\hat{y}_{ij(val)}$.

```
/* Read FTMS_long_cal data into SAS, and create a new
   dataset (ftms_long_cal) for the calibration data */

data ftms_long_cal;
   infile 'FTMS_long_cal.txt';
   input cluster id y lod ybar;
run;

/* Fit the random-intercept censored regression model
   across calibration samples */

proc nlmixed data=ftms_long_cal XTOL=1E-12 method=GAUSS
   qpoints=100;
   parms alpha=0.7 beta=5 sigma2=0.7 tau2=0.5;
   bounds sigma2 tau2 >= 0;
   pi = constant('pi');
   mu = alpha + a_i + beta*ybar;
   if y > lod then
   ll = (1/(sqrt(2*pi*sigma2)))*exp(-(y-mu)**2/(2*sigma2));
   if y <= lod then
   ll = probnorm((y-mu)/sqrt(sigma2));
   L=log(ll);
   model y ~ general(L);
   random a_i ~ normal(0, tau2) subject = id;
   predict mu out=yexp_cal(KEEP = pred);
   predict alpha out=alpha(KEEP = pred);
   predict beta out=beta(KEEP = pred);
   predict sigma2 out=sigma(KEEP = pred);
   predict tau2 out=tau(KEEP = pred);
   by cluster;
run;
```

We can now estimate $\hat{y}_{ij(val)}$, considering α (`alpha`), β (`beta`), σ^2 (`sigma2`) and τ^2 (`tau2`) as fixed (input) variables.

```
/* Read FTMS_long_val data into SAS, and create a new
   dataset (ftms_long_val) for the validation data */

data ftms_long_cal;
   infile 'FTMS_long_cal.txt';
   input cluster id y lod ybar alpha beta sigma2 tau2;
run;
```

```
/* Fit the random-intercept censored regression model
   across validation samples */

proc nlmixed data=ftms_long_val XTOL=1E-12 method=GAUSS
   qpoints=100;
   pi = constant('pi');
   mu = alpha + a_i + beta*ybar;
   if y > lod then
   ll = (1/(sqrt(2*pi*sigma2)))*exp(-(y-mu)**2/(2*sigma2));
   if y <= lod then
   ll = probnorm( (y - mu) / sqrt(sigma2) );
   L=log(ll);
   model y ~ general(L);
   random a_i ~ normal(0, tau2) subject = id;
   predict mu out=yexp_val(KEEP = pred);
   by cluster;
run;
```

Finally, to fit this random-intercept censored regression model based on the CR Re-est approach, we simply have run the same code as for the CR Prep approach, this time separately for the calibration and validation data.

B2. R code for selecting optimal fraction of selected clusters by cross-validation

For the variable selection within calibration approach, presented in Section 4 of this article, we provide R-code based on the penalized R-package (Goeman, 2016). This method requires combined optimization of the fraction of selected clusters F (based on the estimate of the random-intercept variance) and the ridge penalty λ. Once the optimal subset of clusters is selected, the final diagnostic rule can be built based on this subset using the calibration data and can be applied on the validation data to assess the predictive performance.

```
# load the required R-packages
>library('caret')
>library('penalized')

# Read the final calibration data (final estimates of
# average expression - adjusted for the LOD - and estimates
# of random-intercept variance)
>data<-read.table("data_new_cal.txt",sep=",",dec=".")
>data<-t(data)
>tau<-read.table("tau.txt",dec=".")
>tau<-t(tau)
```

```
>tau<-as.matrix(tau)
>group<-read.table("group_cal.txt")
>group<-as.matrix(group-1)

# define k folds
# nfolds<-10 # for k=10-fold
>nfolds<-dim(group)[1] # for loocv
>n<-dim(group)[1]
>fold<-createFolds(1:n, k = nfolds, list = T)

# optimize fraction of selected clusters (quant)
# with respect to error-rate
>quant<-quantile(tau, probs = seq(0, 1, 0.05))
>error<-matrix(NA,length(fold),length(quant)-1)
>lambdagrid<-exp(seq(log(0.1),log(100),
+ by=((log(100)-log(0.1))/50)))

>k<-1
>while (k<=length(fold)){
+  print(k)
+  f<-fold[[k]]
+  data_out<-as.data.frame(data[f,])
+  group_out<-group[f]
+  data_in<-as.data.frame(data[-f,])
+  group_in<-group[-f]
+  lambdaopt<-rep(0,length(quant)-1)
+  j<-1
+  while (j<=length(quant)-1){
+    print(j)
+    data_keep_in<-data_in[,which(tau>=quant[j])]
+    data_keep_out<-as.matrix(data_out[,which(tau>=quant[j])])
+    Dev<-rep(NA,length(lambdagrid))
+    for  (i in 1:length(lambdagrid)){
+      modelfit<-penalized(group_in,data_keep_in,
 +     lambda2=lambdagrid[i])
+      betas<-coefficients(modelfit,"penalized")
+      beta0<-coefficients(modelfit)[1]
+      linpredi<-beta0+data_keep_out%*%betas
+      prob<-exp(linpredi)%*%1/(1+exp(linpredi))
+      loglik<-(group_out*log(prob))+((1-group_out)*log(1-prob))
+      Dev[i]<- -2*sum(loglik)
+    }
```

```
+    lambdaopt[j]<-lambdagrid[Dev==min(Dev)]
+    modelfit<-penalized(group_in,data_keep_in,
+    lambda2=lambdaopt[j])
+    betas<-coefficients(modelfit,"penalized")
+    beta0<-coefficients(modelfit)[1]
+    data_keep_out<-as.matrix(data_out[,which(theta>=quant[j])])
+    linpredi<-beta0+data_keep_out%*%betas
+    prob<-exp(linpredi)%*%1/(1+exp(linpredi))
+    error[k,j]<-length(which(group_out!=(prob>0.5)))
+    /length(group_out)
+    j<-j+1
+  }
+  k<-k+1
+}

# select quant based on cross-validated error-rate
cv_error<-colSums(error)
plot(cv_error)
jopt<-max(which(cv_error==min(cv_error)))
```