Meta-analysis of full ROC curves: Additional flexibility by using semiparametric distributions of diagnostic test values

Annika Hoyer[a*], Oliver Kuss[a]

Abstract

Diagnostic test accuracy studies frequently report on sensitivities and specificities for more than one threshold of the diagnostic test under study. Although it is obvious that the information from all thresholds should be used for a meta-analysis, in practice frequently only a single pair of sensitivity and specificity is selected. To overcome this disadvantage, we recently proposed a statistical model for the meta-analysis of such full receiver operating characteristic (ROC) curves that uses the relationship between a ROC curve and a bivariate model for interval-censored data. In this model, diagnostic tests values reported by the single studies were assumed to follow a parametric distribution. We propose a generalization of this model that allows for a flexible semiparametric modelling of the underlying distribution of the diagnostic test values by using the idea of piecewise constant hazard modelling. We show the results of a simulation study that indicates that the approach works reasonable in practice. Finally, we illustrate the model by the example of population-based screening for type 2 diabetes mellitus.

Keywords: Meta-analysis; ROC curve; piecewise constant model; interval-censored data

1 Introduction

Diagnostic test accuracy studies frequently report on sensitivities and specificities for several different thresholds of the diagnostic test under study. Although it is obvious that the information and observations from all thresholds should be used for a meta-analysis [1, 2], in practice only a single pair of sensitivity and specificity is frequently selected, a practice which is also not discouraged by Cochrane [3].

[a]German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Institute for Biometrics and Epidemiology

*Correspondence to: Annika Hoyer, Deutsches Diabetes-Zentrum, Institut für Biometrie und Epidemiologie, Auf'm Hennekamp 65, 40225 Düsseldorf, Germany. E-Mail: annika.hoyer@ddz.uni-duesseldorf.de

However, methods for the meta-analysis of full receiver operating characteristic (ROC) curves have already been proposed [4, 5, 6, 7, 8, 9, 10, 11, 12, 13], but we noted in previous work [14] that all of them come with at least one disadvantage. For example, some models require that the number of thresholds has to be identical across all studies, some ignore the concrete values of the given thresholds, others use standard inverse-variance two-step approaches where estimation uncertainty from the first step is ignored in the second step. In response to these disadvantages we introduced a model [14] that uses the relationship between a ROC curve and a bivariate time-to-event model for interval-censored data. In this model, diagnostic test values were assumed to follow a parametric distribution, such as Weibull, log-normal or log-logistic. But as already noted in the discussion of that previous paper, a still better fit may be achieved using more irregular distributions with a minimal number of parametric assumptions.

The objective of this article is to show how the well-known idea of piecewise constant hazards, which dates back at least to Holford [15], can be used to fit very flexible distributions of diagnostic test values in the meta-analysis of full ROC curves.

To this task, we first introduce an example data set in Section 2 which will be used for illustrating the new approach. Afterwards in Section 3, we proceed by presenting the statistical methods and a small simulation study in Section 4 which is used to evaluate the piecewise constant model in contrast to the bivariate time-to-event model [14]. In Section 5 we give the results for the motivating example data set and conclude with a critical discussion of the piecewise constant model.

2 Data set

To illustrate our new model, we use a data set from diabetes research which is based on two existing systematic reviews [16, 17]. Both systematic reviews report on glycated haemoglobin A1c (HbA_{1c}, measured in %) for the population-based screening of type 2 diabetes mellitus. Contrary to alternative diagnostic procedures as fasting plasma glucose (FPG) or the oral glucose tolerance test (OGTT) which are based on plasma glucose measurements, HbA_{1c} has some additional advantages as there is, for example, no need for fasting. Furthermore, measuring HbA_{1c} leads to more reliable results which is due to less biologic variability and less pre-analytic instability [18]. The current threshold for diagnosing type 2 diabetes is 6.5 as recommended by the American Diabetes Association (ADA) [19] and the World Health Organization (WHO) [20]. Such threshold values are of considerable relevance as they determine the number of diseased persons and the corresponding health care costs.

For both systematic reviews, only one HbA_{1c} threshold per included single study was selected although most studies report on several of them. Moreover, the authors did not

perform any kind of meta-analysis to get summary estimates for sensitivity, specificity or a summary ROC (SROC) curve. As a consequence, we screened all single studies again and reconstructed the contingency tables for each reported HbA_{1c} threshold. This finally led to 38 included studies which reported 124 pairs of sensitivity and specificity from 26 different thresholds. A standard analysis selecting only one threshold per study as presented in [16, 17] would therefore discard more than 70% of all available observations.

The complete data set was also used in our previous publication where it is given as Supplementary Web Material [14]. Additionally, the data set is available on request from the authors.

3 Methods

In the following, we first give a brief introduction to our previously proposed approach for the meta-analysis of ROC curves which uses a bivariate time-to-event model for interval-censored data [14]. Afterwards, we present our new piecewise constant approach for modelling SROC curves.

3.1 Meta-analysis of full ROC curves based on interval-censored data

The starting point of our previously proposed approach was the fundamental insight that summary receiver operating characteristic (SROC) curves can be estimated using a bivariate time-to-event model for interval-censored data [14]. To be concrete, we first note that diagnostic test values can be considered as interval-censored. This is because each study reports on a complete contingency table with the numbers of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) tested participants for each given threshold. As we do not have individual participant data available but only aggregated data, we do not know the exact diagnostic test value of each study participant. We only have information on the number of participants with diagnostic test values that lie above or below a pre-specified diagnostic threshold. Therefore, we are able to construct intervals, where the lower and upper bounds are defined by the diagnostic thresholds, which include the diagnostic test values of the participants, but we only know the numbers of diagnostic test values between the two bounds and not their exact values. Based on this assumption, we proposed to model the log-transformed diagnostic test values in the population of diseased and non-diseased using three kinds of distributions (exemplarily illustrated for the population of diseased D^+):

- the Weibull distribution with density

$$f(y_{D^+};\mu_{D^+},\phi_{D^+}) = \frac{\phi_{D^+} y_{D^+}^{\phi_{D^+}-1}\exp(-(y_{D^+}/\mu_{D^+})^{\phi_{D^+}})}{\mu_{D^+}^{\phi_{D^+}}}, \tag{1}$$

- the log-normal distribution with density

$$f(y_{D^+};\mu_{D^+},\phi_{D^+}) = \frac{\exp(-[\log(y_{D^+})-\mu_{D^+}]^2/(2\phi_{D^+}))}{y_{D^+}\sqrt{2\pi\phi_{D^+}}}, \text{and} \tag{2}$$

- the log-logistic distribution with density

$$f(y_{D^+};\mu_{D^+},\phi_{D^+}) = \frac{\pi\exp(-\pi[\log(y_{D^+})-\mu_{D^+}]/(\sqrt{3}\phi_{D^+}))}{y_{D^+}\sqrt{3}\phi_{D^+}(1+\exp(-\pi[\log(y_{D^+})-\mu_{D^+}]/(\sqrt{3}\phi_{D^+})))^2}, \tag{3}$$

  where μ_{D^+} and ϕ_{D^+} represents location and scale parameters, respectively, and y_{D^+} the diagnostic test values in the population of diseased.

In the framework of time-to-event models, the events of interest are being test-positive or test-negative in the population of diseased and non-diseased, respectively. Furthermore, the diagnostic test values y_{D^+} and y_{D^-} are treated as the "time variable", meaning the event probabilities sensitivity and specificity are decreasing or increasing with rising diagnostic test values. Finally, the outcome is log-transformed which leads to an accelerated failure time (AFT) model, going along with the advantage of a unified linear predictor where a random effect can be added.

The final model is specified as follows:

$$\log(y_{D^-}) = b_{D^-} + \epsilon_{D^-} + u_{iD^-}, \tag{4}$$

$$\log(y_{D^+}) = b_{D^+} + \epsilon_{D^+} + u_{iD^+}, \tag{5}$$

with

$$\begin{pmatrix} u_{iD^-} \\ u_{iD^+} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{D^-} & \rho\sigma_{D^-}\sigma_{D^+} \\ \rho\sigma_{D^-}\sigma_{D^+} & \sigma^2_{D^+} \end{pmatrix}\right], \tag{6}$$

where y_{D^+} and y_{D^-} are the diagnostic test values in the population of diseased and non-diseased, respectively. u_{iD^+} and u_{iD^-} are the study-specific (indexed by i, $i \in 1,\ldots,N$) random effects with their corresponding variances $\sigma^2_{D^+}$ and $\sigma^2_{D^-}$ and correlation parameter ρ which are used to model potential across study correlations. Furthermore, b_{D^+} and b_{D^-} are the location parameters after log-transforming y_{D^+} and y_{D^-} and therefore transformations of the original location and scale parameters μ_{D^+}, ϕ_{D^+}, μ_{D^-} and ϕ_{D^-} of the Weibull, log-normal and log-logistic distribution. Moreover, ν_{D^+} and ν_{D^-} denote the corresponding scale parameter after the log-transformation. For example, in the population of diseased, we get [14]

- for the Weibull distribution: $b_{D^+} = \log(\mu_{D^+})$ and $\nu_{D^+} = 1/\phi_{D^+}$,
- for the log-normal distribution: $b_{D^+} = \mu_{D^+}$ and $\nu_{D^+} = \phi_{D^+}$, and
- for the log-logistic distribution: $b_{D^+} = -\phi_{D^+} \log(\mu_{D^+})$ and $\nu_{D^+} = 1/\phi_{D^+}$.

ϵ_{D^+} and ϵ_{D^-} denote the error terms and determine the residual distributions after log-transforming the diagnostic test values y_{D^+} and y_{D^-}. These are the Gumbel, normal and logistic distribution that correspond to the Weibull, log-normal and log-logistic distribution, respectively.

Finally, the corresponding survival functions are used to predict the sensitivities and specificities at various thresholds.

3.2 The piecewise constant model

As a flexible alternative to the bivariate time-to-event model introduced in Section 3.1, we propose to use a semiparametric piecewise constant model for interval-censored data to estimate SROC curves. There is plenty room for discussion if the piecewise constant model is truly a "semiparametric" model. We are aware of descriptions of this model as "parametric", "weakly parametric" "quasi-semiparametric", "semiparametric", or "nonparametric", but decided to follow Ibrahim et al. [24] to use the term "semiparametric" in the work reported here. For a more comprehensive introduction to piecewise constant models we refer to [21, 22, 23].

Let T denote the diagnostic test value which is interval-censored as we do not know the exact test value of a participant but only the interval $I = (L, R]$ in which it lies. L is the last threshold prior to R where the event (being test positive) is absent. If a participant has no event, i.e. is never tested positive, L is the last threshold which denotes the absence of the disease and R is set to be missing. In case of our piecewise constant model, we partition the continuous diagnostic test values into J intervals with thresholds $0 = \tau_0 < \tau_1 < ... < \tau_J = \infty$, where J is the number of different thresholds across all single studies included in the meta-analysis. The j-th interval is defined as $(\tau_{j-1}, \tau_j]$. As piecewise constant models are in general defined in terms of hazards, we assume that the baseline hazard in the population of non-diseased is constant within each interval:

$$\lambda_{D^-}(t) = \lambda_{jD^-} \quad \text{for } t \text{ in } (\tau_{j-1}, \tau_j]. \tag{7}$$

That means, the baseline hazard $\lambda_{jD^-}(t)$ is modelled using J parameters $\lambda_1, ..., \lambda_J$.

Analogously to (7), a piecewise constant baseline hazard for the population of diseased can be assumed. Both piecewise constant modelled baseline hazards could then be linked by a bivariate random effect to model potential heterogeneities. However, we propose to use

a univariate approach based on a proportional hazards model [25] to limit the number of parameters that have to be estimated.

In terms of the proportional hazards model, we assume the baseline hazard to be piecewise constant, leading to the model:

$$\lambda_{ij}(t) = \lambda_{jD^-}(t) \exp(x_i^T \beta), \tag{8}$$

where $\lambda_{ij}(t)$ is the hazard of individual (study) i in interval j and $\lambda_{jD^-}(t)$ is the piecewise constant baseline hazard for interval j. The cumulative hazard function is then defined as [21]

$$\Lambda_i(t) = \sum_{j=1}^{J} \lambda_{jD^-}(t) \exp(x_i^T \beta) \max(0, \min(\tau_j - \tau_{j-1}, t - \tau_{j-1})), \tag{9}$$

and the survival function can be written as

$$S_i(t) = \exp(-\Lambda_i(t)). \tag{10}$$

The log likelihood function corresponding to the model can be formulated as

$$l = \sum_{i=1}^{n} k(\delta_i(\log(S_i(L) - S_i(R))) + (1 - \delta_i)(\log(S_i(L)))), \tag{11}$$

where δ_i denotes the censoring indicator for study i which is 1 for subjects with events that are interval censored and 0 in case of right-censored data. To account for the different study and group (within study) sizes, we additionally include k as the number of events in the population of diseased and non-diseased, respectively.

As we assume the baseline hazard in the population of non-diseased to be piecewise constant, the population of diseased participants is modelled by simply including a binary covariate x_D which is 1 in case the considered disease is present and 0 when the disease is absent. Additionally, we add a univariate random effect to model potential heterogeneities. The final model can be written as

$$\lambda_{ij}(t) = \lambda_{jD^-}(t) \exp(\beta_D x_D + u_i) \tag{12}$$

with $u_i \sim N(0, \sigma_u^2)$.

4 Simulation study

To compare the performance of the piecewise constant model to our previously proposed approach for the meta-analysis of full ROC curves [14], we conducted a simulation study. This approach was chosen as competitor because it compensates the disadvantages of other models and is also based on interval-censored data. The complete simulation program was written in SAS 9.3 (SAS Institute Inc., Cary, NC, USA).

4.1 Setting

In order to mirror realistic parameter constellations, we varied our model parameters according to the systematic reviews on diabetes screening [16, 17]. As true models for data generation we used the Weibull and the log-logistic model because univariate Weibull models fulfil the proportional hazard assumption whereas univariate log-logistic models correspond to a proportional odds model [26]. As diagnostic test values have to be simulated for the population of diseased as well as for the non-diseased, we chose true parameter values for the Weibull distribution to guarantee proportional hazards between both groups. In line with [14] we varied the true parameters of both models in order to achieve true areas under the curve (AUC) of 65% and 85%. Additionally, we varied the correlation between the random effects (0, 0.28 or 0.85) which illustrates the absence as well as moderate and strong potential heterogeneity across studies.

Table 1 shows the true values for the sensitivity and specificities for the simulation study.

PLACE TABLE 1 APPROXIMATELY HERE

4.2 Data generation

After varying and combining our model parameters, we get in total 24 different scenarios. For each of them, 1.000 meta-analyses were simulated. Analogous to [14], the number of studies per meta-analysis was drawn from a uniform distribution, varying between 10 and 30. Also the number of participants per study follows a uniform distribution, ranging from 30 to 300. Furthermore, the number of thresholds per study was generated from a uniform distribution between 1 and 4 while rounding to the nearest integer. Depending on the number of thresholds, the actual threshold values were simulated as given in [14] and also in Table 2.

PLACE TABLE 2 APPROXIMATELY HERE

To finally generate the numbers of diseased and non-diseased, a prevalence was drawn from a uniform distribution, ranging from 0.3 to 0.5.

To arrive at diagnostic test values for each study participant, these values were randomly drawn from the true underlying distribution, i.e. the Weibull or log-logistic distribution. The algorithm which is used to generate these numbers from the Weibull and log-logistic distribution is presented by Hoyer et al. [14]. To be concrete, a bivariate random effect with predefined values for σ_{D^+}, σ_{D^-} and ρ according to Equation (6) was first generated. With regard to Equation (4) and (5), the generated random effect was added to b_{D^-} and b_{D^+}. Together with predefined values for ν_{D^-} and ν_{D^+}, we got Gumbel and logistic distributed

diagnostic test values on the log-sale. Back-transforming these values to the original scale yielded the diagnostic test values of each participant. Finally, the simulated diagnostic test value of each participant was compared to the previously defined threshold to declare whether people were tested true positive or true negative. Based upon this, the complete contingency tables for each study and each threshold could be determined.

4.3 Estimation methods

For each simulated meta-analysis, we used the Weibull, log-logistic and the piecewise constant model to estimate SROC curves. For all of them, we used PROC NLMIXED for parameter estimation and Gaussian quadrature with its default options to enable a fair comparison between the different approaches. The number of pieces and their concrete values for the piecewise constant model was chosen to be equal to the number of different thresholds per meta-analysis. That means, we are not restricted to set the number of pieces to a fixed value for all simulated meta-analysis but using instead a flexible number. Starting values for the bivariate time-to-event models were obtained as described in [14] using PROC LIFEREG. To be concrete, we fitted two univariate models in PROC LIFEREG to achieve starting values for the fixed effects parameters $b_{D^-}, b_{D^+}, \nu_{D^-}$ and ν_{D^+}. Furthermore, starting values for the random effects parameters $\sigma_{D^-}, \sigma_{D^+}$ and ρ were calculated as raw estimates from the two univariate fits. In order to identify starting values for each of the pieces, we fitted a univariate Weibull model for interval-censored data using PROC LIFEREG including a binary covariate indicating the disease status. We used the estimated Weibull shape and scale parameters, μ_w and ν_{wd}, where d denotes the disease status, to calculate the Weibull hazard rate h_w using $h_{wm} = \nu_{wd}^{-\mu_w} \mu_w t_m^{\mu_w - 1}$ where t_m are the different threshold values per meta-analysis. To avoid specifying a BOUNDS statement, we first log-transformed the h_{wm} which were back-transformed in PROC NLMIXED, that is, we used $\log(h_{wm})$ as starting values for the pieces. To obtain starting values for β_D we transformed the estimated regression coefficient for the binary covariate β_w from PROC LIFEREG using $\log(HR) = \beta_D = -\beta_w * k_w$, thus using the relation between the AFT and the proportional hazard interpretation of parameters from the Weibull model. As the the random effect variance σ_u^2 corresponds to the variability of the $\log(HR) = \beta_D$ between the single studies, we fitted univariate Weibull models for each study again using PROC LIFEREG. Afterwards, we used PROC MEANS to calculate the variance of these $\log(HR)$ as starting value for σ_u^2. It has to be noted that fitting an univariate Weibull model per study did not work in many cases which is due to non-convergence. In this case, we decided to set the starting value for the random effect variance σ_u^2 to the fixed value of 0.5 which can be seen as a conservative choice.

4.4 Results

In the following paragraphs, we give a brief overview of the results of our simulation study, focussing on the estimated sensitivities and specificities at various thresholds. The scenarios with a true AUC of 0.85 and a true correlation of $\rho = 0.28$ are exemplarily presented in Tables 3 and 4. The remaining results are given as Supplementary Web Material. Additionally, Table 5 shows the complete simulation results addressing numerical robustness.

Bias All parametric time-to-event models perform best when the estimated model is also the true underlying model. As expected, the piecewise constant model performs worse compared to the parametric models in terms of bias in case the true underlying parametric distribution is known. However, when the Weibull model which is also a proportional hazard model serves as true underlying model, the piecewise constant model performs slightly better compared to scenarios where data are generated from the log-logistic model. This is to be expected because the piecewise constant model also requires a proportional hazard assumption which is not fulfilled when the log-logistic approach as proportional odds model is the true model. Especially for the highest evaluated threshold of 7.0, a large bias of the piecewise constant model is observed. In general, sensitivity is overestimated for lower thresholds and underestimated for higher thresholds (specificity vice versa). In particular, estimates for lower thresholds from the piecewise constant model are less biased and only worse than the respective true underlying parametric model. There is also no clear dependence on the correlation structure visible.

PLACE TABLE 3 APPROXIMATELY HERE

Empirical coverage In terms of empirical coverage (to the 95% level), the respective results from the piecewise constant model are in most cases below 95%. In line with the results concerning bias, the parametric models perform best and almost always better than the piecewise constant model in case the true underlying distribution coincide with the estimated model. The performance of the piecewise constant model is even worse when the proportional hazards assumption is violated, i.e. in case data were generated from the log-logistic model. Again, no relation between coverage and correlation structure is visible. The lowest coverages of the piecewise constant model are observed for lower thresholds. A possible reason may be small estimated variances, resulting in narrow confidence intervals as it was also described in [14]. In contrast to the parametric models, the coverage of the piecewise constant model is not decreasing with rising thresholds.

PLACE TABLE 4 APPROXIMATELY HERE

Convergence To address the numerical performance of the different models, we report on the number of converged simulation runs (with a maximum of 1,000). These numbers are satisfying for all models as in every setting more than 900 converged runs are reached. As the piecewise constant model is more complex than the parametric models (which is due to a higher number of estimated parameters), its performance is in some scenarios inferior to the Weibull and log-logistic model.

PLACE TABLE 5 APPROXIMATELY HERE

5 Example

To illustrate the practical applicability of the piecewise constant model, we use the data set introduced in Section 2. As this data set reports on 26 different HbA_{1c} thresholds, we also used 26 pieces which are defined by them. To compare our new model to existing ones, we included the results of our previous proposed approach [14]. The complete numerical results for various thresholds are given in Table 6. Figure 1 displays the estimated SROC curves of the piecewise constant model and the parametric time-to-event models. Moreover, the underlying ROC curves of the single studies are included as lightgrey lines and dots. The SROC curve of the piecewise constant model lies below the SROC curve of the log-logistic model, but agrees well with the SROC curve of the Weibull model. While the estimated specificities of the piecewise constant model suit well with the parametric time-to-event models with slightly larger differences for the lowest threshold, there are more discrepancies between the estimated sensitivities at moderate thresholds. For example, at a threshold of 6.0 the piecewise constant model estimates a sensitivity of 56% compared to sensitivities of nearly 60% using the parametric time-to-event models.

At an HbA_{1c} value of 6.5 which is the current threshold for diagnosing type 2 diabetes recommended by the ADA [19] and WHO [20], the piecewise constant model leads to a sensitivity of 33.7% and a specificity of 98.5%. This corresponds to a Youden index of 0.322. In line with the results presented in [14], we found that 5.9 is the threshold with the maximum Youden index of 0.461 meaning that the optimal HbA_{1c} threshold should be chosen equal to 5.9 instead of 6.5, at least if one is willing to give equal weights to sensitivity and specificity.

To calculate the AUC and its 95% confidence interval, we used the trapezoidal rule and a nonparametric bootstrap approach. For the piecewise constant model, we estimated an AUC of 78.8% [75.6%; 81.8%] which is slightly lower compared to the AUC of the log-logistic model (83.7% [77.0%; 90.3%]), but in line with the AUC of the Weibull model (76.3% [72.4%; 84.1%]).

The SAS code used to fit the piecewise constant model for this example can be found in the Supporting Web Materials.

PLACE TABLE 6 APPROXIMATELY HERE

PLACE FIGURE 1 APPROXIMATELY HERE

6 Discussion

In this article, we proposed a novel model for the meta-analysis of ROC curves that incorporates all observations and information from the single studies. Contrary to existing approaches that also use these information [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14], we suggest the usage of a semiparametric piecewise constant model which does not require the parametric specification of underlying distributions, but estimates this distribution explicitly from the data. As such, it avoids prespecifying parametric distributions for the diagnostic test values as in most cases no medical rationale for such a choice is given. As we propose to use all reported different threshold values for defining the pieces, the model accounts for all given diagnostic thresholds and returns the corresponding estimated meta-sensitivities and meta-specificities for displaying a full summary ROC curve. Of course, it also possible to set the number of pieces to a predefined value which could be lower than the number of different thresholds. However, this leads to reduced flexibility as the estimated SROC curve is evaluated at less thresholds. Assuming a proportional hazard within each piece for the diseased and non-diseased participants, only a single random effect is required to model potential heterogeneity and correlations between sensitivity and specificity across studies. It is important to note that estimating only a single semiparametric hazard does not imply that only data from the non-diseased group is used.

In line with the quadrivariate meta regression approach proposed by Hoyer and Kuss [13], the piecewise constant model can also be used to compare the SROC curves of two (or more) different diagnostic tests. Moreover, it is also possible to include other covariates to compare diagnostic accuracy studies in a meta-regressive sense. Similar to our parametric time-to-event model for interval-censored data [14], the piecewise constant model can also be used if individual participant data, that is data for each participant with the exact observed diagnostic test value, are available.

In summary, our model compensates for the disadvantages of existing approaches and, moreover, does not need specifications of the underlying distributions of the diagnostic test values in the population of diseased and non-diseased. This can be seen as an additional advantage over our previously proposed parametric model [14].

In a small simulation study we showed that in case distributions of diagnostic test values in the population of diseased and non-diseased are known, parametric approaches perform better and should be the models of choice, but the piecewise constant model might be a valid alternative, especially when nothing is known about the possible distribution of the diagnostic test values, e.g. for irregular mixture distributions. However, in our simulation with only mild deviations from the parametric assumptions, the models using parametric distributions were robust enough.

Of course, our model comes with some disadvantages. Although the piecewise constant approach is of a semiparametric nature, there is still a proportional hazard assumption required. This might be a restriction for real data sets where this assumption is not fulfilled. However, in our simulation study we observed moderate biased estimates when the input data were not generated from a proportional hazards model. As an potential alternative that avoids the explicit assumption of proportional hazards, it would be possible to estimate two separate piecewise constant hazard functions, one for the group of non-diseased and one for the group of diseased participants. Unless one restricts the numbers of pieces this would essentially double the number of parameters which comes with the risk of convergence problems. Moreover, we are not able to give a model selection criteria for comparing our model to the parametric time-to-event model for interval-censored data [14] which is due to different likelihood functions. But of course it is possible to compare piecewise constant models using a different number of pieces in terms of AIC and BIC to each other.

For future work we plan to assess versions of the semiparametric piecewise constant model that do not need a proportional hazard assumption, but model separate piecewise constant hazards for the groups of diseased and non-diseased, linking them by a bivariate random effect.

References

[1] Sutton AJ, Higgins JP. Recent developments in meta-analysis. *Statistics in Medicine* 2008;**27**(5):625-650.

[2] Trikalinos TA, Balion CM, Coleman CI, et al. Chapter 8: Meta-analysis of Test Performance When There is a 'Gold Standard'. *J Gen Intern Med.* 2012;**27:**Supplement:56-66. doi: 10.1007/s11606-012-2029-1

[3] Macaskill P, Gatsonis C, Deeks JJ, et al. Chapter 10: analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C (eds) *Cochrane handbook for systematic reviews of diagnostic test accuracy version 1.0.* The Cochrane Collaboration; 2010. Available at:

http://methods.cochrane.org/sdt/sites/methods.cochrane.org.sdt/files/uploads/Chapter%2010%20-%20Version%201.0.pdf (accessed 24 September 2018)

[4] Hamza TH, Arends LR, van Houwelingen HC, Stijnen T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Med Res Methodol.* 2009;**9:**73. doi: 10.1186/1471-2288-9-73

[5] Kester AD, Buntinx F. Meta-analysis of ROC curves. *Med Decis Making.* 2000;**20:**430–439.

[6] Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics.* 2003;**59:**936–946.

[7] Poon WY. A latent normal distribution model for analysing ordinal responses with applications in meta-analysis. *Stat Med.* 2004;**23,**2155–2172.

[8] Bipat S, Zwinderman AH, Bossuyt PM, Stoker J. Multivariate random-effects approach: for meta-analysis of cancer staging studies. *Acad Radiol.* 2007;**14:**974-984.

[9] Putter H, Fiocco M, Stijnen T. Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. *Biom J.* 2010;**52:**95–110. doi: 10.1002/bimj.200900073

[10] Riley RD, Takwoingi Y, Trikalinos T, et al. Meta-Analysis of Test Accuracy Studies with Multiple and Missing Thresholds: A Multivariate-Normal Model. *Journal Biom Biostat.* 2014;**5:**3.

[11] Martínez-Camblor P. Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Stat Method Med Res.* 2017;**26:** 5–20. doi: 10.1177/0962280214537047

[12] Steinhauser S, Schumacher M, Rücker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Med Res Methodol.* 2016;**16:**97. doi: 10.1186/s12874-016-0196-1

[13] Hoyer A, Kuss O. Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model. *Stat Methods Med Res.* 2016. pii: 0962280216661587. (Epub ahead of print)

[14] Hoyer A, Hirt S, Kuss O. Meta-analysis of full ROC curves using bivariate time-to-event models for interval-censored data. *Research Synthesis Methods.* 2017. doi: 10.1002/jrsm.1273 [Epub ahead of print]

[15] Holford TR. Life tables with concomitant information. *Biometrics.* 1976;**32:**587–597.

[16] Bennett CM, Guo M, Dharmage SC. HbA(1c) as a screening tool for detection of Type 2 diabetes: a systematic review. *Diabet Med.* 2007;**24:**333–343.

[17] Kodama S, Horikawa C, Fujihara K, et al. Use of high-normal levels of haemoglobin A_1C and fasting plasma glucose for diabetes screening and for prediction: a meta-analysis. *Diabetes Metab Res Rev.* 2013;**29:**680–692. doi: 10.1002/dmrr.2445

[18] International Expert Committee. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. *Diabetes Care.* 2009;**32:**1327-1334. doi: 10.2337/dc09-9033

[19] American Diabetes Association. Classification and diagnosis of diabetes. Sec. 2. In Standards of Medical Care in Diabetes. *Diabetes Care.* 2015;**38**(Suppl. 1):S8–S16. doi: 10.2337/dc16-er09

[20] World Health Organization. Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. Abbreviated report of a WHO consultation. Geneva, Switzerland, World Health Organization; 2011.

[21] Gong Q, Fang L. Comparison of different parametric proportional hazards models for interval-censored data: a simulation study. *Contemp Clin Trials.* 2013;**36:**276–283.

[22] Demarqui FN, Loschi RH, Dey DK, Colosimo EA. A class of dynamic piecewise exponential models with random time grid. *Journal of Statistical Planning and Inference.* 2012;**142:**728–742.

[23] Friedman M. Piecewise Exponential Models for Survival Data with Covariates. *The Annals of Statistics.* 1982;**10:**101–113.

[24] Ibrahim JG, Chen MH, Sinha D. Bayesian Survival Analysis. New York: Springer-Verlag; 2001.

[25] Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological).* 1972;**34:**187–220.

[26] Doksum KA, Gasko M. On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review/Revue Internationale de Statistique.* 1990;**58:**243–252

Figure 1: Estimated summary ROC curves of the piecewise-constant and the time-to-event models. Light grey solid lines and dots depict the estimated sensitivities, specificities and ROC curves of the single studies

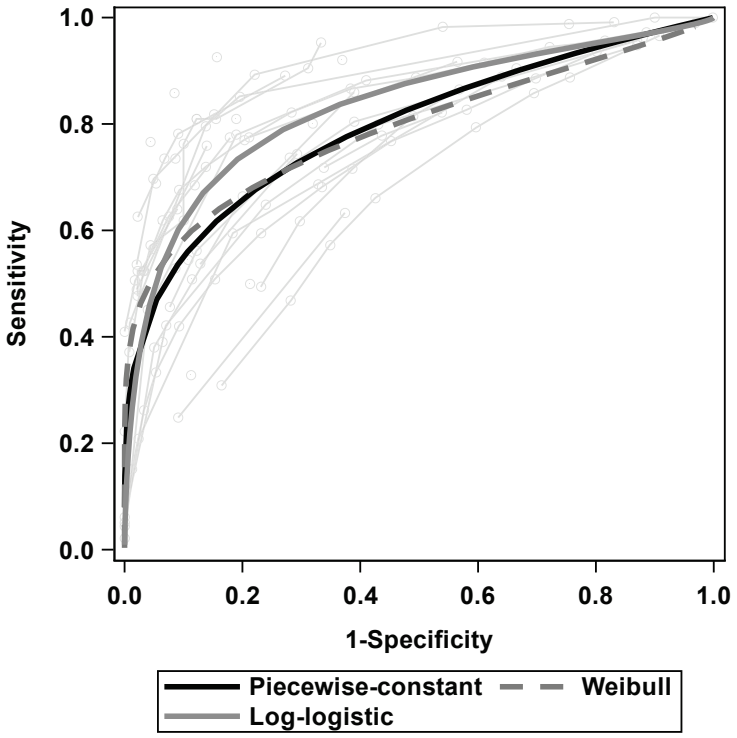

Table 1: True sensitivities and specificities used for the simulation study. True values for the Weibull, log-normal and log-logistic model were also presented in [14]

Model	Threshold	Sensitivity (AUC = 0.65)	Sensitivity (AUC = 0.85)	Specificity
Weibull	5.0	93%	98%	14%
	5.5	74%	92%	45%
	6.0	36%	75%	88%
	6.5	4%	41%	100%
	7.0	0%	7%	100%
Log-logistic	5.0	95%	99%	12%
	5.5	75%	95%	48%
	6.0	35%	76%	84%
	6.5	10%	39%	96%
	7.0	2%	12%	100%

Table 2: Generated thresholds per study according to [14]

Number of thresholds	Simulated distribution
1	U(6.0, 6.5)
2	U(5.6, 6.0), U(6.4, 6.8)
3	U(5.4, 5.8), U(6.1, 6.5), U(6.7, 7.1)
4	U(5.3, 5.7), U(5.8, 6.2), U(6.3, 6.7), U(6.8, 7.2)

Table 3: Bias (in percentage points): sensitivity (sens) and specificity (spec)

True model/ AUC/ Correlation	Threshold	Estimated model					
		Piecewise constant		Weibull		Log-logistic	
		Sens	Spec	Sens	Spec	Sens	Spec
Weibull/ 0.85/0.28	5.0	-1.3	2.0	-0.1	0.4	1.1	-4.8
	5.5	-3.6	1.9	-0.5	0.5	1.6	5.8
	6.0	-5.8	-3.5	-1.0	-0.7	-2.6	1.9
	6.5	-0.3	-1.2	-0.5	-0.2	-4.9	-1.5
	7.0	21.7	-12.7	1.4	0.0	4.3	-0.3
Log-logistic/ 0.85/0.28	5.0	-2.8	1.1	-2.1	11.9	-0.1	0.6
	5.5	-9.0	-3.5	-4.0	1.8	-0.6	0.2
	6.0	-9.3	-5.8	-1.0	-4.6	-1.0	-0.4
	6.5	2.6	-0.2	6.9	0.3	0.6	-0.2
	7.0	21.8	-14.2	2.1	0.8	1.0	0.0

Table 4: Empirical coverage (in %): sensitivity (sens) and specificity (spec)

True model/ AUC/ Correlation	Threshold	Estimated model					
		Piecewise constant		Weibull		Log-logistic	
		Sens	Spec	Sens	Spec	Sens	Spec
Weibull/ 0.85/0.28	5.0	78.0	83.0	93.6	93.5	34.2	58.6
	5.5	61.1	74.3	93.8	92.9	77.9	90.5
	6.0	67.3	76.5	93.6	90.4	94.2	80.8
	6.5	75.8	92.4	92.9	81.7	85.9	63.7
	7.0	53.1	76.7	86.5	100	92.1	58.3
Log-logistic/ 0.85/0.28	5.0	40.8	70.0	46.6	16.7	93.4	93.2
	5.5	11.1	69.7	79.1	85.5	92.2	92.9
	6.0	49.6	77.7	90.8	87.6	91.7	93.1
	6.5	75.0	80.9	88.1	88.4	92.2	92.7
	7.0	55.1	30.4	93.0	7.6	92.7	92.0

Table 5: Number of converged runs

	Estimated model					
	Piecewise constant		Weibull		Log-logistic	
True model/ Correlation	AUC 0.65	AUC 0.85	AUC 0.65	AUC 0.85	AUC 0.65	AUC 0.85
Weibull/0.00	920	944	998	1000	971	988
Weibull/0.28	907	938	998	999	990	991
Weibull/0.85	904	927	988	992	965	985
Log-logistic/0.00	986	990	994	990	995	996
Log-logistic/0.28	985	988	993	998	998	997
Log-logistic/0.85	990	984	983	987	986	991

Table 6: Results from the different models for the diabetes data set

Threshold	Estimated model					
	Piecewise constant		Weibull		Log-logistic	
	Sens [95%-CI]	Spec [95%-CI]	Sens [95%-CI]	Spec [95%-CI]	Sens [95%-CI]	Spec [95%-CI]
5.0	97.2 [96.5; 97.8]	10.4 [8.1; 12.9]	89.5 [85.6; 93.3]	27.9 [14.5; 41.3]	97.5 [95.9; 99.0]	8.9 [5.3; 12.5]
5.5	82.6 [78.8; 86.3]	52.2 [43.8; 60.5]	78.1 [70.8; 85.4]	58.5 [37.8; 79.3]	87.7 [80.9; 94.4]	52.2 [41.2; 63.2]
6.0	56.0 [48.3; 63.7]	89.3 83.6; 94.9]	59.9 [48.5; 71.3]	88.6 [74.6; 100]	60.3 [45.5; 75.1]	90.8 [87.2; 94.5]
6.5	33.7 [25.0; 42.5]	98.5 [97.0; 100]	36.8 [23.0; 50.5]	99.3 [97.4; 100]	26.9 [14.7; 39.1]	98.7 [98.1; 99.3]
7.0	23.6 [15.1; 32.1]	99.6 [99.1; 100]	15.5 [4.6; 26.5]	100 [100;100]	9.0 [3.9; 14.1]	99.8 [99.7; 99.9]