Meta-analysis of full ROC curves with flexible parametric distributions of diagnostic test values

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

Abstract

Diagnostic accuracy studies often evaluate diagnostic tests at several threshold values, aiming to make recommendations on optimal thresholds for use in practice. Methods for meta-analysis of full receiver operating characteristic (ROC) curves have been proposed, but still have deficiencies. We recently proposed a parametric approach that is based on bivariate time-to-event models for interval-censored data to this task. To increase the flexibility of that approach and to address the open point of model selection, we here suggest to use the generalized F family of distributions which includes currently used distributions for the bivariate time-to-event model as special cases. The results of a simulation study are given as well as an illustration by an example of population-based screening for type 2 diabetes mellitus.

Keywords: Meta-analysis; ROC curve; time-to-event model; interval-censored data; Generalized F distribution

1 Introduction

In medical research, diagnostic accuracy studies often evaluate diagnostic tests at several threshold values, aiming to make recommendations on optimal thresholds for use in practice. In the past, methods for the meta-analysis of diagnostic studies frequently focussed on summarizing only a selected single pair of sensitivity and specificity from each study which leads to a waste of the majority of observations. Methods for the meta-analysis of full receiver operating characteristic (ROC) have been proposed [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], however, they come with several disadvantages [11]. These are, for example, the need for an identical number of thresholds across single studies, ignoring concrete threshold values or

[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

applying two-step approaches that partly neglect estimation uncertainty from the first step in the second step. As an alternative and to overcome these disadvantages, we proposed an approach based on bivariate time-to-event models for interval-censored data [11]. Thereby, we assume that the underlying diagnostic test values in the population of diseased and non-diseased follow parametric distributions, such as the Weibull, log-normal or log-logistic distribution. While a wide variety of distributions is available, we also noted that it is not possible to choose among them using well-defined model selection criteria.

In this article, we extend the model of Hoyer et al. [11] by referring to the family of generalized F (GF) distributions, which is governed by 4 parameters and includes all of our previously used distributions as special cases. Extending this basic set of distributions by one additional parameter yields the generalized log-logistic, Burr III and Burr XII distribution as further flexible special cases of the GF distribution. As a consequence, we are now able to apply model selection criteria as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) to choose the most appropriate model because all investigated distributions are members of the GF family.

The article is organized as follows: first, we introduce our motivating example data set in Section 2. Afterwards, we describe the used statistical methods in Section 3 which are compared in a simulation study as presented in Section 4. In Section 5, the results for the example data set are presented. Finally, in Section 6 we conclude with a critical discussion of our approach.

2 Data set

Our work is motivated by two existing systematic reviews on population-based screening for type 2 diabetes mellitus [12, 13]. Both of them investigate the diagnostic performance of glycated haemoglobin A1c (HbA_{1c}, measured in %) as a marker for type 2 diabetes. Contrary to existing tests for diagnosing type 2 diabetes as fasting plasma glucose (FPG) and the oral glucose tolerance test (OGTT) which are based on glucose measurement, HbA_{1c} is increasingly preferred due to several advantages. For example, there is no need for fasting, less biologic variability and less pre-analytic instability of HbA_{1c} [14]. The American Diabetes Association (ADA) [15] and the World Health Organization (WHO) [16] recommend 6.5% as the optimal threshold value for declaring a test positive.

Both systematic reviews come with two drawbacks: First, they do not report on meta-analytic summary measures as, for example, sensitivity and specificity, and second, the authors selected only one threshold per single study although in most cases at least two are reported. Especially the second point is of importance because it indicates that many observations are lost. Therefore, we screened all single studies again, leading to in total

38 studies reporting on 124 pairs of sensitivity and specificity at 26 different diagnostic thresholds. A meta-analysis based on only one selected threshold per single study would therefore discard more than 70% of the available observations. Table 1 shows exemplarily for the first two studies how the data set look like. The complete data set was also used in [11] where it is given in the supplementary material. Alternatively, it can be requested from the authors.

PLACE TABLE 1 APPROXIMATELY HERE

3 Statistical methods

Hereinafter, we briefly introduce the already published bivariate time-to-event model for estimating summary receiver operating (SROC) curves [11]. Afterwards, alternative parametrizations of that model using the generalized F family of distributions are presented.

3.1 Bivariate time-to-event model

Our recent approach for the meta-analysis of full ROC curves is based on a bivariate time-to-event model for interval-censored data [11].

The central assumption is that the diagnostic test values are interval-censored as we only know if (and how many of) these lie above or below the given thresholds of the respective study. As we are interested in sensitivity and specificity, we have to keep diseased and non-diseased participants apart and finally arrive at the close relation between a ROC curve and a bivariate time-to-event model for interval-censored data. For modelling diagnostic test values in the two populations of diseased and non-diseased we used three different distributions which are in the following given for the population of diseased (D^+), and are defined analogously for the non-diseased (D^-) [11]:

- 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

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

- 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.

Explicitly naming the parallels of our model to the class of time-to-event models, our events of interest when meta-analysing full ROC curves, are being test positive or test negative in the population of diseased and non-diseased, respectively. Additionally, we consider the diagnostic test values as the time scale. Sensitivity is then the event probability in the population of diseased, (or 1-specificity in the population of non-diseased) which can be interpreted, with the diagnostic test values on the x-axis, as a cohort life-table estimator where sensitivity decreases with increasing thresholds [11]. Moreover and to finally be in the context of accelerated failure time (AFT) models, we log-transformed the outcome, thus the diagnostic test values. The advantage of that transformation is that a unified linear predictor corresponding to our assumed distributions is achieved where random effects can unequivocally be added.

The model equation then reads as:

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

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

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^-} denote the diagnostic test values in the population of diseased and non-diseased, respectively. The study-specific (indexed by i, $i \in 1,\ldots,N$) random effects u_{iD^+} and u_{iD^-} with their corresponding variances $\sigma^2_{D^+}$ and $\sigma^2_{D^-}$ and correlation parameter ρ are used to model potential heterogeneity and across study correlations. b_{D^+} and b_{D^-} are location parameters after suitable 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. Additionally, λ_{D^+} and λ_{D^-} denote the scale parameters after log-transformation and will be used explicitly for predicting sensitivities and specificities as presented in Section 3.3. As given in [11], we get (exemplary for the population of diseased)

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

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

Finally, ϵ_{D^+} and ϵ_{D^-} are the corresponding error terms which define the distributions after the log-transformation of y_{D^+} and y_{D^-}. These distributions are the Gumbel, normal and logistic distribution in case the Weibull, log-normal and log-logistic distributions are used on the original scale, respectively.

3.2 The generalized F distribution

The bivariate time-to-event model from Section 3.1 comes with the need of pre-specifying a parametric distribution for the underlying diagnostic test values. This could be considered as a disadvantage, because external information on the type of distribution is rarely available. Moreover, Hoyer et al. [11] do not report on model selection criteria to distinguish between the Weibull, log-normal and log-logistic model. To overcome these disadvantages, we propose to use the GF family of distributions which includes the Weibull, log-normal and log-logistic distribution as special cases and additionally comes along with two special features. Firstly, we are much more flexible regarding possible distributions of the underlying diagnostic test values because the GF distribution has 4 parameters as compared to our previously used distributions with only 2 parameters. Secondly, as all distributions are members of the same family, classical model selection criteria as AIC or BIC are applicable to compare the fit of the different models.

We use the parametrization for the GF family as proposed by Prentice and Cox [17, 18]. The probability density function (pdf) for the population of diseased D^+ is

$$f_{GF}(y_{D^+}) = \frac{\delta_{D^+} \exp(-b_{D^+} m_{1D^+}/\lambda_{D^+}) y_{D^+}^{(\delta_{D^+}/\lambda_{D^+})m_{1D^+}} (m_{1D^+}/m_{2D^+})^{m_{1D^+}}}{y_{D^+} \delta_{D^+} B(m_{1D^+}, m_{2D^+})[1 + (m_{1D^+}/m_{2D^+})(\exp(-b_{D^+})y_{D^+})^{\delta_{D^+}/\lambda_{D^+}}]^{(m_{1D^+}+m_{2D^+})}}, \tag{7}$$

where $B(m_{1D^+}, m_{2D^+})$ is the beta function evaluated at m_{1D^+} and m_{2D^+}. The parameters m_{1D^+}, m_{2D^+} and δ_{D^+} are defined as follows:

$$m_{1D^+} = 2\left(q_{D^+}^2 + 2p_{D^+} + q_{D^+}(q_{D^+}^2 + 2p_{D^+})^{1/2}\right)^{-1}, \tag{8}$$

$$m_{2D^+} = 2\left(q_{D^+}^2 + 2p_{D^+} - q_{D^+}(q_{D^+}^2 + 2p_{D^+})^{1/2}\right)^{-1} \tag{9}$$

and

$$\delta_{D^+} = \left(m_{1D^+}^{-1} + m_{2D^+}^{-1}\right)^{1/2} = \left(q_{D^+}^2 + 2p_{D^+}\right)^{1/2}, \tag{10}$$

with $-\infty < q_{D^+} < \infty$ and $p_{D^+} > 0$. Analogously, the pdf can be defined for the population of non-diseased D^- with the respective parameters $b_{D^-}, \lambda_{D^-}, p_{D^-}$ and q_{D^-}.

The GF distribution is thus characterized by the 4 parameters $p_\bullet, q_\bullet, b_\bullet, \lambda_\bullet$ and is therefore very flexible. Thereby, $b_\bullet$ and $\lambda_\bullet$ represent the location and scale parameter, respectively, and $m_{1\bullet}$ and $m_{2\bullet}$ as transformations of $p_\bullet$ and $q_\bullet$ denote the, potentially non-integer, degrees of freedom. Moreover, by restricting $p_\bullet$ and $q_\bullet$ to specific values, the number of parameters can be reduced, resulting in further well-known distributions as, for example, the generalized gamma, the generalized log-logistic, the Burr III and Burr XII as well as the Weibull, log-normal and log-logistic distribution. Figure 1 shows the relationships between the members of the GF family. Having experienced serious numerical problems with the generalized gamma distribution, we focus in the following on the GF, the generalized log-logistic, the Burr III and Burr XII, the Weibull, log-normal and log-logistic distribution.

PLACE FIGURE 1 APPROXIMATELY HERE

3.3 Predicting sensitivities and specificities

As we are not explicitly interested in the actual parameters of the different underlying distributions, but in sensitivities and specificities at various thresholds, we use the best linear unbiased prediction (BLUP) principle to predict them at different thresholds using the corresponding survival functions. Focussing on sensitivity and estimating the parameters in the population of diseased (D^+), the survival functions for the Weibull, log-normal and log-logistic distributions were already presented in [11]:

- for the Weibull distribution,

$$Sensitivity = \exp(-(y \exp(-b_{D^+}))^{\frac{1}{\lambda_{D^+}}}, \tag{11}$$

- for the log-normal distribution,

$$Sensitivity = 1 - \Phi\left(\frac{1}{\lambda_{D^+}}(\log(y) - b_{D^+})\right), \tag{12}$$

  where Φ is the cdf of the normal distribution, and

- for the log-logistic distribution,

$$Sensitivity = 1/\left(1 + \exp\left(\frac{-b_{D^+}}{\lambda_{D^+}}\right) y^{\frac{1}{\lambda_{D^+}}}\right). \tag{13}$$

For the remaining members of the GF family we obtain [18]:

- for the generalized log-logistic distribution,

$$Sensitivity = 1 - F_{Beta}\left(x_{GLL}, a_{GLL}, b_{GLL}\right), \tag{14}$$

where F_{Beta} is the cumulative distribution function (cdf) of the beta distribution with $x_{GLL} = \frac{1/p_{D^+}}{1/p_{D^+}+1/p_{D^+}\exp(\sqrt{2p_{D^+}}(\log(y)-b_{D^+})/\lambda_{D^+})}$ and $a_{GLL} = b_{GLL} = 1/p_{D^+}$,

- for the Burr III distribution,

$$Sensitivity = 1 - F_{Beta}\left(x_{BIII}, a_{BIII}, b_{BIII}\right), \tag{15}$$

where $x_{BIII} = \frac{1}{1+m_{1D^+}\exp(\delta_{D^+}(\log(y)-b_{D^+})/\lambda_{D^+})}$, $a_{BIII} = 1$, $b_{BIII} = m_{1D^+}$, $q_{D^+} = -(1 - p_{D^+})\sqrt{\frac{2}{2-p_{D^+}}}$, and $m_{1D^+} = \frac{2}{q_{D^+}^2+2p_{D^+}+q_{D^+}\sqrt{q_{D^+}^2+2p_{D^+}}}$,

- for the Burr XII distribution,

$$Sensitivity = 1 - F_{Beta}\left(x_{BXII}, a_{BXII}, b_{BXII}\right), \tag{16}$$

where $x_{BXII} = \frac{m_{2D^+}}{m_{2D^+}+\exp(\delta_{D^+}(\log(y)-b_{D^+})/\lambda_{D^+})}$, $a_{BXII} = m_{2D^+}$, $b_{BXII} = 1$, the parameters q_{D^+} and m_{2D^+} being defined as $q_{D^+} = (1 - p_{D^+})\sqrt{\frac{2}{2-p_{D^+}}}$ and $m_{2D^+} = \frac{2}{q_{D^+}^2+2p_{D^+}-q_{D^+}\sqrt{q_{D^+}^2+2p_{D^+}}}$,

- for the generalized F distribution:

$$Sensitivity = 1 - F_{Beta}\left(x_{GF}, a_{GF}, b_{GF}\right), \tag{17}$$

where $x_{GF} = \frac{m_{2D^+}}{m_{2D^+}+m_{1D^+}\exp(\delta_{D^+}(\log(y)-b_{D^+})/\lambda_{D^+})}$, $a_{GF} = m_{2D^+}$, $b_{GF} = m_{2D^+}$ with $m_{1D^+} = 2\left(q_{D^+}^2 + 2p_{D^+} + q_{D^+}(q_{D^+}^2 + 2p_{D^+})^{1/2}\right)^{-1}$ and $m_{2D^+} = 2\left(q_{D^+}^2 + 2p_{D^+} - q_{D^+}(q_{D^+}^2 + 2p_{D^+})^{1/2}\right)^{-1}$.

4 Simulation

To compare the model using the generalized log-logistic, Burr III, Burr XII and GF distribution to the previously used Weibull, log-normal and log-logistic approach [11], we conducted a simulation study. The corresponding simulation program was written in SAS 9.4 (SAS Institute Inc., Cary, NC, USA).

4.1 Setting

For our simulation study, simulated data were drawn from the Weibull, log-normal, log-logistic and GF distribution. That means, these 4 distributions served as our 'true underlying distributions'. We used the estimates of $b_{D^+}, b_{D^-}, \lambda_{D^+}$ and λ_{D^-} for the Weibull, log-normal and log-logistic distribution, as well as in addition the estimates of $p_{D^+}, p_{D^-}, q_{D^+}$ and q_{D^-} for the GF distribution from the diabetes example in Section 2 as true parameter values for our simulation. That means, the simulation settings mirror the situation from the diabetes data set. In case of the GF distribution, the true parameter values were chosen in a way that only the GF, but not the generalized log-logistic, the Burr III, or the Burr XII model was the true model. Additionally, we varied the true correlation between the random effects (by setting it to 0, 0.28 or 0.85) in order to mimic zero, moderate and strong heterogeneity across studies. In Table 2, the true sensitivities and specificities used for our simulation study are depicted.

PLACE TABLE 2 APPROXIMATELY HERE

4.2 Data generation

Combining all parameter settings led to 12 different simulation scenarios. We simulated 1.000 meta-analyses for each of them. In line with [11], the number of studies for each meta-analysis was generated from a uniform distribution, ranging from 10 to 30. The number of participants per study was also drawn from a uniform distribution and varied between 30 and 300. Numbers of diseased participants were generated based on a uniformly distributed prevalence between 0.3 and 0.5. For the number of thresholds per study, we also used a uniform distribution to generate values between 1 and 4 which were rounded to the nearest integer. Analogous to [11], the actual threshold values were simulated from a uniform distribution where the range depended on the number of thresholds. To be concrete, for example, a study with two different thresholds, will report on a first threshold between 5.6 and 6.0 and a second threshold between 6.4 and 6.8, whereas the first threshold from a study reporting on three thresholds, varies between 5.4 and 5.8, the second one between 6.1 and 6.5 and the third one between 6.7 and 7.1.

The exact diagnostic test values for every study participant were simulated from the true underlying distribution, that means from the Weibull, log-normal, log-logistic or the GF distribution. In case of the Weibull, log-normal and log-logistic distribution we proceeded as proposed in [11]. That means, we first generated a bivariate normally distributed random effect according to Equation (6) with predefined $\sigma^2_{D^+}, \sigma^2_{D^-}$, and ρ. With respect to Equations (4) and (5), this random effect was added to b_{D^-} and b_{D^+}. In combination with previously defined λ_{D^-} and λ_{D^+}, Gumbel, normal and logistic distributed diagnostic test values on

the log-scale were generated. Back-transforming led to diagnostic test values of each study participant on the original scale for the Weibull, log-normal and log-logistic distribution. These values were finally compared with the generated thresholds aiming to classify true positives and true negatives as well as the resulting fourfold tables.

For generating random numbers from the GF distribution, we adapted the algorithm implemented in the *flexsurv* package of the statistical software R [19] to SAS. This algorithm is based on the representation of the GF family as given in [18] and inverse transform sampling.

4.3 Estimation methods and outcomes

For each of the 1.000 meta-analyses per simulation scenario, we estimated the Weibull, log-normal, log-logistic, generalized log-logistic, Burr III, Burr XII and the GF model to predict sensitivities and specificities at different thresholds. SAS PROC NLMIXED with Gaussian quadrature and its default options was used for parameter estimation. Starting values for this procedure were determined using SAS PROC LIFEREG in line with [11] to get starting values for the fixed as well as for the random effects parameters. We used SAS PROC LIFEREG assuming a log-logistic distribution to obtain starting values for the generalized log-logistic, Burr III, Burr XII and the GF model because this is a sub-model of the other approaches. Moreover, starting values for p and q were set to 1 in the population of diseased, whereas q was set to 0 in the population of non-diseased in case the Weibull, log-normal or log-logistic distribution served as true model. If the GF distribution was the true underlying distribution, starting values for p were set to 1.8 and 0.5 in the population of diseased and non-diseased, respectively. Starting values for q were assumed to be 0.5 in the population of diseased and -0.6 in the population of non-diseased. These choices were inspired by estimates for the diabetes example data set.

Sensitivities and specificities were evaluated at predefined thresholds (5.0, 5.5, 6.0, 6.5 and 7.0), as well as their corresponding t-confidence intervals with degrees of freedom determined by the default option in SAS PROC NLMIXED.

We focused on bias, empirical coverage, number of converged simulation runs and the chosen model as our outcomes of interest. For model choice we used AIC and BIC, defined as follow:

$$AIC = 2f(\hat{\theta}) + 2k$$

and

$$BIC = 2f(\hat{\theta}) + k\log(s),$$

where f is the negative of the marginal log-likelihood, $\hat{\theta}$ is the vector of parameter estimates, k is the number of parameters and s is the number of subjects, here the number of studies.

4.4 Results

In the following, we report on the results of our simulation study which is focussed on the predicted sensitivities and specificities (as our main parameters of interest). Results for the scenario with a true underlying correlation of 0.28 and the GF distribution as true model are presented in Tables 3 and 4. All remaining simulation results can be found in the Supplementary Web Materials. Tables 5 and 6 show the complete results concerning numerical robustness and model selection.

Bias In terms of bias, the Weibull, log-normal and log-logistic model performed best in case they were also the true underlying model. However, also the more complex models (i.e. generalized log-logistic, Burr III, Burr XII and GF) led to satisfying results that are mostly in magnitude of the log-normal and log-logistic model or even better. Especially in case the GF distribution was the true model, simpler models, as Weibull, log-normal and log-logistic, performed slightly worse. However, the maximum amount of bias of the GF model was 8.3 when data were generated from the log-normal distribution. This was higher compared to maximum bias of the log-normal or log-logistic model with 5.4 and 6.8, respectively. Remarkably, the Burr III model led to worst results of all models whereas in particular the Burr XII and the GF model showed very low biases for all evaluated thresholds. In most cases, the different models overestimated sensitivities for lower thresholds, but underestimated them for higher ones (specificity vice versa). No obvious dependence on the true underlying correlation structure was visible.

PLACE TABLE 3 APPROXIMATELY HERE

Empirical coverage Concerning empirical coverage (to the 95% level), the results of all models were rather satisfying but in many cases below the expected 95%. In line with bias, the best results were obtained if the estimated model was also the true underlying model. Especially the Weibull model performed worse in case another model was used for simulating the input data set. The generalized log-logistic, Burr XII and GF model performed well and without huge outliers regardless of the true underlying distribution. Moreover, these models led to superior results compared to the simpler ones if the GF distribution served as true model. This was observed especially for the lowest and highest evaluated threshold. While the Burr III model performed well in case the Weibull, log-normal or log-logistic distribution were the true underlying models, results for the GF distribution were rather unsatisfactory, especially for higher thresholds where coverages of 23.2% were observed. Again, the underlying correlation structure had no impact on the results.

PLACE TABLE 4 APPROXIMATELY HERE

Number of converged runs In terms of numerical robustness, we report on the total number of converged simulation runs (out of 1.000) for each model. These results were extremely satisfying, especially for the log-normal and log-logistic model where for many settings the maximum number of converged simulation runs were observed. For the Burr XII and the generalized F model, slightly worse results were obtained. But this was expected as these models go along with more and more complexity resulting in an increased numerical instability. Worst results were observed for the Burr III model in case the GF distribution served as true model.

PLACE TABLE 5 APPROXIMATELY HERE

Model selection criteria In Table 6, we report on the number that each model has been selected due to the minimal AIC or BIC. The maximum that can be achieved is therefore 1.000. For the Weibull, log-normal and log-logistic model it is obvious that they were the models of choice in case they served also as the true distribution of the diagnostic test values. Settings based on the Weibull distribution and a correlation of 0.28 or 0.85 led to the result that in the majority of cases the Weibull or the log-normal model was selected. For all remaining settings, the Weibull model was rarely chosen. If the log-normal model was the true underlying distribution, also the Burr III and Burr XII were frequently selected. As expected, models with a higher number of parameters, namely the Burr XII and the GF model, were preferred when data were generated from the GF distribution. Again, the Burr III model showed worst results in case the GF distribution was the true model.

PLACE TABLE 6 APPROXIMATELY HERE

5 Example

To illustrate the practical application of our model, we use the example of population-based screening for type 2 diabetes presented in Section 2. Focussing on the estimated sensitivities and specificities at various thresholds as the parameters with most practical relevance, the corresponding results are given in the Supplementary Web Materials. Additionally, Figure 2 depicts the different estimated SROC curves as well as the ROC curves from the underlying single studies. All SROC curves are generated based on predicted sensitivities and specificities at various thresholds. As the predicted sensitivities and specificities from the log-logistic, generalized log-logistic, Burr III, Burr XII and GF model are nearly identical, SROC curves from these models are indistinguishable. However, with respect to model selection criteria (shown in Table 7) the model based on the GF distribution, which is actually the most complex one, has the lowest AIC and BIC and is therefore the model of choice for the diabetes

example. In Table 7 we also report on the estimated values for $m_{1D^+}, m_{1D^-}, m_{2D^+}$ and m_{2D^-} for the models that contain these parameters. It can be seen that all of them differ from 1 (albeit sometimes with large confidence intervals), indicating that the underlying distribution of the diagnostic test values in the diabetes data set differs from the Weibull, log-normal or log-logistic distribution. Consequently, more flexible models as the GF model reflect the underlying distribution in a better way. The predicted sensitivities and specificities from the Weibull model differ, especially for lowest and highest thresholds, from the results of the other models which is due to the left-skewness of the distribution.

PLACE TABLE 7 APPROXIMATELY HERE

At the currently recommended HbA_{1c} threshold of 6.5 [15, 16], all models lead to high specificities between 98% and 99% and low sensitivities of about 31%. As we aim to evaluate whether HbA_{1c} is an appropriate primary screening test for type 2 diabetes, we are more interested in higher sensitivities which is associated with a lower diagnostic threshold. Moreover, even in case sensitivity and specificity are weighted equally (which corresponds to a choice based on the Youden index), the optimal threshold would be chosen between 5.5 and 6.0 which is quite lower than in current guidelines.

Finally, we estimated the areas under the curve (AUC) for each model using the trapezoidal rule and a nonparametric resampling bootstrap for the 95% confidence intervals (Table 7). Reflecting the virtually indistinguishable SROC curves for the models with more than 2 parameters, also their AUCs are largely identical.

In the Supplementary Web Materials, we present a SAS macro that can be used to fit the generalized log-logistic, Burr III, Burr XII and the GF model.

PLACE FIGURE 2 APPROXIMATELY HERE

6 Discussion

In this article, we proposed an extension of the already published bivariate time-to-event model for interval-censored data which can be used to estimate SROC curves [11]. This approach allows us to use information on each and every threshold reported by the single studies included in a systematic review. In contrast to existing approaches which can also use this information [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], we suggest to model the distribution of the underlying diagnostic test values by using the flexible GF family of distributions. This family is characterized by 4 parameters and includes the Weibull, log-normal, log-logistic, generalized log-logistic, Burr III and Burr XII distribution as special cases.

Therefore, our suggested model can cope with distributions which are more flexible than the Weibull, log-normal and log-logistic used previously. Second, we are now able to apply

model selection criteria as the AIC or the BIC to choose the most appropriate model. Note that a likelihood ratio test would be insufficient for model selection here as not all models are nested.

Of course, all other advantages going along with the model of Hoyer et al. [11] are still present. These include, for example, that the number of thresholds as well as their concrete values can be different across studies, or that the exact threshold values are used in modelling.

In a small simulation study, we showed that the generalized log-logistic, Burr III, Burr XII and GF model led to satisfying results that are comparable to the Weibull, log-normal and log-logistic approach and in some cases even better.

Of course, our model has some limitations. Because the GF distribution is very flexible, it might be difficult to estimate the 4 parameters in case only small data sets are available. The still existing necessity of pre-specifying the distribution of the diagnostic test values might be considered another disadvantage of our parametric approach. It would be a definite advantage to have an alternative nonparametric approach available, for example for more irregular or mixtures of distributions. Moreover, the estimated distributions of the diagnostic test values are only marginal ones and distributions of the diagnostic test values in the single studies may differ from the overall one. Finally, we proposed to model the diagnostic test values on a log-scale. Depending on the diagnostic test or biomarker under evaluation, it could be meaningful to forgo this transformation and to use other distributions as the normal or logistic distribution (as for example reported by Steinhauser et al. [9]) instead of members of the GF family of distributions. However, especially the GF distribution might be a flexible approach that can cope with skewed as well as with non-skewed distributions, neglecting this slight disadvantage.

In summary, bivariate parametric models for interval-censored data based on the GF family of distributions are a very flexible approach for the meta-analysis of full ROC curves.

References

[1] 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

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

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

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

[5] 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.

[6] 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

[7] 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.

[8] 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

[9] 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

[10] 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.* 2018;27(5):1410-1421

[11] 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.* 2018;9(1):62-72

[12] 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.

[13] 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

[14] 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

[15] 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

[16] 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.

[17] Prentice RL. Discrimination among some parametric models. *Biometrika.* 1975;**62:**607–614.

[18] Cox C. The generalized F distribution: An umbrella for parametric survival analysis. *Stat Med.* 2008;**27:**4301–4312.

[19] Jackson CH. flexsurv: a Platform for Parametric Survival Modeling in R. *Journal of Statistical Software.* 2016;**70:**1–33.

Figure 1: The family of generalized F distributions. The term $q(p)$ in the transition from the generalized F to the Burr distributions is defined as $q(p) = (1-p)[2/(2-p)]^{1/2}$ with $p < 2$

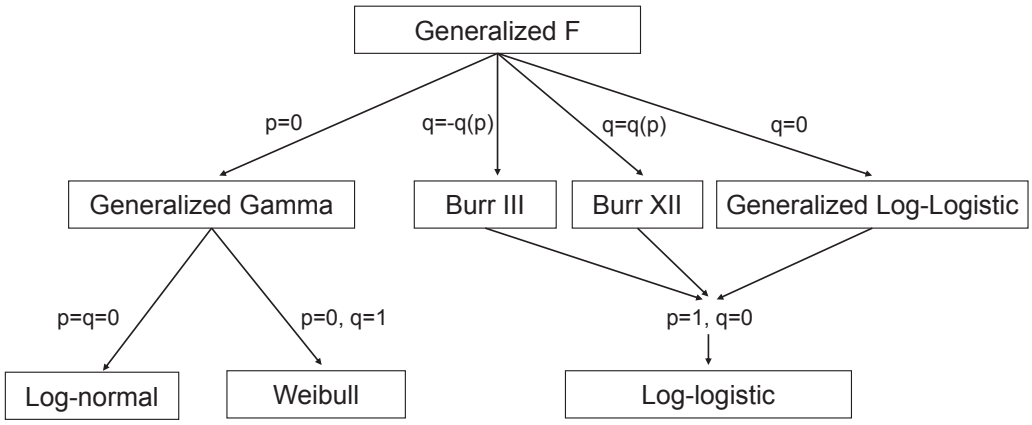

Figure 2: Estimated summary ROC curves of the different time-to-event models. Light grey solid lines and dots depict the estimated sensitivities, specificities and ROC curves of the single studies

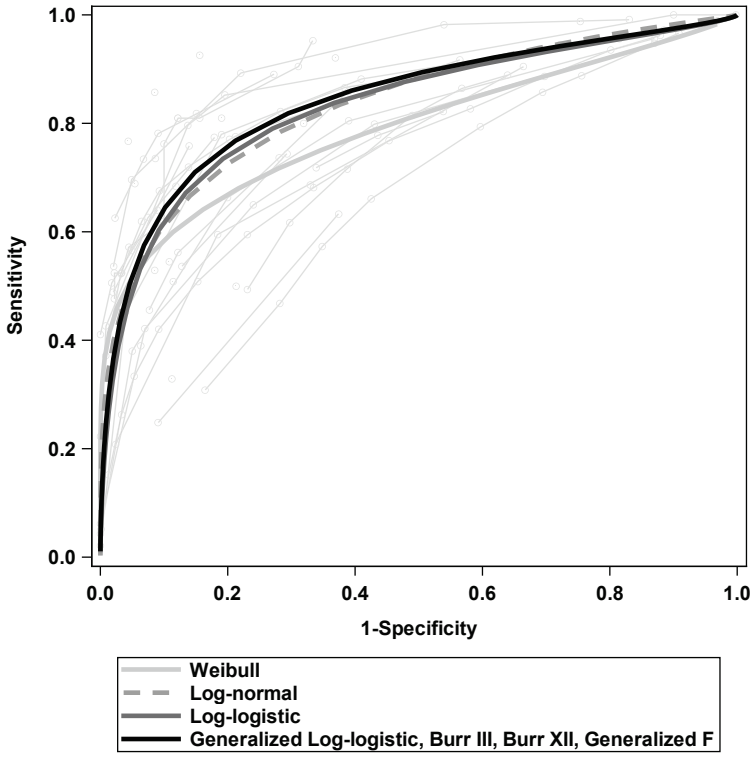

Table 1: First two studies reported in the systematic reviews of Bennett et al. [12] and Kodama et al. [13], bold entries are the originally included thresholds

Study	Threshold HbA_{1c}	TP	FN	FP	TN
Badings et al.	**5.9**	**574**	**262**	**682**	**1389**
Choi et al.	5.0	617	18	7735	1005
	5.1	607	28	7123	1617
	5.2	600	35	6302	2438
	5.3	581	54	5331	3409
	5.4	563	72	4318	4422
	5.5	550	85	3356	5384
	5.6	522	113	2473	6267
	5.7	**489**	**146**	**1774**	**6966**
	5.8	457	178	1206	7534
	5.9	429	206	813	7927
	6.0	393	242	568	8172
	6.2	332	303	280	8460
	6.6	236	399	70	8670

Table 2: True sensitivities and specificities for simulation

Model	Threshold	Sensitivity	Specificity
Weibull	5.0	96%	16%
	5.5	87%	50%
	6.0	65%	91%
	6.5	31%	100%
	7.0	5%	100%
Log-normal	5.0	98%	10%
	5.5	88%	53%
	6.0	61%	91%
	6.5	30%	99%
	7.0	10%	100%
Log-logistic	5.0	98%	7%
	5.5	89%	47%
	6.0	65%	89%
	6.5	33%	98%
	7.0	13%	100%
Generalized F	5.0	97%	6%
	5.5	92%	44%
	6.0	80%	81%
	6.5	58%	94%
	7.0	30%	98%

Table 3: Bias (in percentage points): sensitivity (sens) and specificity (spec), bold entries indicate the smallest bias

True model/		Estimated model							
Correlation/	Threshold	Weibull		Log-normal		Log-logistic		Generalized log-logistic	
		Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
Generalized F/0.28	5.0	-0.4	14.8	1.9	5.0	1.2	3.6	0.6	3.5
	5.5	-1.0	1.3	1.2	-0.5	1.1	-1.5	0.6	-1.8
	6.0	-1.4	-4.0	-2.5	-1.0	-1.2	**0.7**	-0.9	1.2
	6.5	0.4	1.9	-3.5	1.7	-2.9	1.6	-2.9	1.6
	7.0	2.7	1.7	1.9	1.3	1.1	0.8	**-0.2**	0.7

True model/		Estimated model					
Correlation/	Threshold	Burr III		Burr XII		Generalized F	
		Sens	Spec	Sens	Spec	Sens	Spec
Generalized F/0.28	5.0	-5.7	3.3	0.6	1.1	**-0.1**	**0.3**
	5.5	-10.9	9.0	0.5	-1.5	**-0.3**	**-0.4**
	6.0	-19.0	6.2	**-0.1**	1.2	-0.7	1.1
	6.5	-25.2	2.6	**0.3**	**1.0**	-0.7	1.1
	7.0	-18.2	1.0	2.8	**0.4**	0.5	0.6

Table 4: Empirical coverage (in %): sensitivity (sens) and specificity (spec), bold entries indicate coverages between 93.6 and 96.4% (95% Wald confidence interval for a binomial proportion of 950 successes out of 1000 trials)

True model/		Estimated model							
Correlation/	Threshold	Weibull		Log-normal		Log-logistic		Generalized log-logistic	
		Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
Generalized F/0.28	5.0	92.3	4.9	27.1	81.7	50.1	90.8	79.2	85.5
	5.5	92.4	82.5	85.4	89.4	84.7	89.3	88.9	84.6
	6.0	93.2	89.1	**95.6**	**93.9**	**94.4**	90.2	92.2	85.1
	6.5	92.1	74.2	91.2	71.1	91.8	68.7	87.7	67.4
	7.0	92.6	2.3	89.5	11.9	90.0	34.3	86.4	50.0

True model/		Estimated model					
Correlation/	Threshold	Burr III		Burr XII		Generalized F	
		Sens	Spec	Sens	Spec	Sens	Spec
Generalized F/0.28	5.0	100	100	81.0	86.8	92.1	83.2
	5.5	100	100	89.8	91.1	94.5	90.4
	6.0	50.0	66.7	**93.6**	88.8	93.5	89.2
	6.5	0.0	33.3	92.7	78.0	91.5	75.9
	7.0	0.0	20.0	92.7	71.1	91.8	59.2

Table 5: Number of converged simulation runs, bold entries indicate the maximum number

True model/	Estimated model						
Correlation/	Weibull	Log-normal	Log-logistic	Generalized log-logistic	Burr III	Burr XII	Generalized F
Weibull/0.00	999	**1000**	**1000**	998	956	761	644
Weibull/0.28	999	**1000**	**1000**	998	956	761	644
Weibull/0.85	999	**1000**	**1000**	998	956	761	644
Log-normal/0.00	**1000**	**1000**	**1000**	998	956	761	644
Log-normal/0.28	**1000**	**1000**	**1000**	998	956	761	644
Log-normal/0.85	**1000**	**1000**	**1000**	998	956	761	644
Log-logistic/0.00	921	998	**1000**	960	997	**1000**	966
Log-logistic/0.28	919	999	**1000**	966	995	997	959
Log-logistic/0.85	864	997	994	922	989	**996**	954
Generalized F/0.00	943	**1000**	**1000**	783	8	653	871
Generalized F/0.28	948	999	**1000**	748	7	709	877
Generalized F/0.85	910	993	**994**	693	3	703	902

Table 6: Model selection: number of selected models (in terms of AIC and BIC) from 1000 meta-analyses, bold entries indicate the maximum number

True model/	Estimated model							
Correlation/	Weibull		Log-normal		Log-logistic		Generalized log-logistic	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Weibull/0.00	**934**	**972**	0	0	0	0	0	0
Weibull/0.28	**473**	**490**	469	469	22	22	0	0
Weibull/0.85	**473**	**490**	469	469	22	22	0	0
Log-normal/0.00	4	4	**444**	**445**	35	35	5	5
Log-normal/0.28	1	1	**465**	**465**	27	27	5	5
Log-normal/0.85	1	1	**465**	**465**	27	27	5	5
Log-logistic/0.00	0	0	82	92	**711**	**829**	51	14
Log-logistic/0.28	0	0	86	93	**735**	**852**	59	16
Log-logistic/0.85	0	0	91	98	**722**	**831**	49	15
Generalized F/0.00	2	4	3	3	127	227	60	42
Generalized F/0.28	0	0	2	3	119	208	43	43
Generalized F/0.85	6	6	3	3	114	189	35	32

True model/	Estimated model					
Correlation/	Burr III		Burr XII		Generalized F	
	AIC	BIC	AIC	BIC	AIC	BIC
Weibull/0.00	36	16	26	11	4	1
Weibull/0.28	21	12	13	7	2	0
Weibull/0.85	21	12	13	7	2	0
Log-normal/0.00	180	204	257	262	75	45
Log-normal/0.28	173	192	259	263	70	47
Log-normal/0.85	173	192	259	263	70	47
Log-logistic/0.00	81	41	65	24	10	0
Log-logistic/0.28	69	28	43	11	8	0
Log-logistic/0.85	65	28	61	26	10	0
Generalized F/0.00	2	2	368	**415**	**438**	307
Generalized F/0.28	2	2	**419**	**461**	415	283
Generalized F/0.85	3	2	397	**453**	**441**	314

Table 7: Diabetes data set: AIC, BIC, model parameters and estimated AUCs of the different models

Estimated model	AIC	BIC	m_{1D^+} [95% CI]	m_{1D^-} [95% CI]	m_{2D^+} [95% CI]	m_{2D^-} [95% CI]	AUC [95% CI]
Weibull	219,093	219,104	-	-	-	-	76.3%
			-	-	-	-	[72.4%; 84.1%]
Log-normal	209,632	209,644	-	-	-	-	83.9%
			-	-	-	-	[78.7%; 86.8%]
Log-logistic	209,356	209,367	-	-	-	-	83.7%
			-	-	-	-	[77.0%; 90.3%]
Generalized log-logistic	209,333	209,348	0.97	1.22	-	-	84.0%
			[0.49; 1.45]	[1.03; 1.40]	-	-	[75.3%; 85.4%]
Burr III	209,315	209,330	0.81	1.16	-	-	83.9%
			[0.65; 0.98]	[1.08; 1.23]	-	-	[82.3%; 85.2%]
Burr XII	209,325	209,340	-	-	1.45	0.93	83.8%
			-	-	[0.96; 1.93]	[0.88; 0.98]	[81.9%; 85.0%]
Generalized F	209,296	209,314	1.43	1.97	2.42	1.52	83.9%
			[-0.17; 3.02]	[0.66; 3.27]	[-1.79; 6.63]	[0.83; 2.21]	[82.3%; 84.7%]