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

Annika Hoyer1 | Oliver Kuss1,2

1German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Institute for Biometrics and Epidemiology, Düsseldorf, Germany
2Medical Faculty, Heinrich Heine University Düsseldorf, Institute of Medical Statistics, Düsseldorf, Germany

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

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, to cover a wide range of distributions of diagnostic test values and to address the open point of model selection, we here suggest to use the generalized $F$ family of distributions that includes previously 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
generalized $F$ distribution, interval-censored data, meta-analysis, ROC curve, time-to-event model

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 focused on summarizing only a selected single pair of sensitivity and specificity from each study that leads to a waste of the majority of observations. Methods for the meta-analysis of full receiver operating characteristic (ROC) curves have been proposed, however, they come with several disadvantages. These are, for example, the need for an identical number of thresholds across single studies, ignoring concrete threshold values or 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. 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 our previous model by referring to the family of generalized $F$ (GF) distributions, which is governed by four 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 distributions 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. In applications, individual participant data (i.e., the concrete test values of each study participant) are rarely...
available, and it is thus difficult to foresee the true underlying distributions of the diagnostic test values. As such, the flexibility of the extended model is necessary to cover a wide range of possible distributions.

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 hemoglobin A1c (HbA1c, 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, HbA1c is increasingly preferred because of several advantages. For example, there is no need for fasting, less biologic variability, and less preanalytic instability of HbA1c.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 looks like. The complete data set was also used in Hoyer et al11 where it is given in the Supporting Information. Alternatively, it can be requested from the authors.

3 | STATISTICAL METHODS

Hereinafter, we briefly introduce the already published bivariate time-to-event model for estimating summary receiver operating characteristic (SROC) curves.11 Afterwards, an extension of that model using the generalized F family of distributions is 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

| Study          | Threshold HbA1c | 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   |
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 correspondence between an ROC curve and a bivariate time-to-event model for interval-censored data. For modeling 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

$$f(y_{D^+}; \mu_{D^+}, \phi_{D^+}) = \mu_{D^+} \phi_{D^+} (\mu_{D^+} y_{D^+})^{\phi_{D^+} - 1} \exp\left(-(\mu_{D^+} y_{D^+})^{\phi_{D^+}}\right),$$  

(1)

- the log-normal distribution with density

$$f(y_{D^+}; \mu_{D^+}, \phi_{D^+}) = \frac{1}{\phi_{D^+} y_{D^+}} \phi(\log(y_{D^+}) - \mu_{D^+}) \phi_{D^+},$$  

(2)

where $\phi$ denotes the probability density function (pdf) of the normal distribution, and

- the log-logistic distribution with density

$$f(y_{D^+}; \mu_{D^+}, \phi_{D^+}) = \frac{\mu_{D^+} y_{D^+}^{\phi_{D^+} - 1}}{(1 + \mu_{D^+} y_{D^+}^{\phi_{D^+}})^2},$$  

(3)

where $\mu_{D^+}$ and $\phi_{D^+}$ represent location and scale parameters, respectively, and $y_{D^+}$ the diagnostic test values in the population of diseased.

Explicitly naming the analogy of our model to the class of time-to-event models, 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 changes (in general, decreases) with increasing thresholds.11 In other words, each individual study participant contributes three pieces of information: its true disease status, the interval in which its diagnostic test value lies (and which is measured on our underlying “time” scale), and finally, the study from which it originates.

As we have to include random intercepts into the linear predictors to account for correlated test values within studies, and as it is not immediately obvious where to add them, we log-transform our outcome, that is, the diagnostic test values. After log-transformation, we are in context of accelerated failure time (AFT) models, which are defined by a unified linear predictor corresponding to our assumed distributions where random effects can unequivocally be added.17

The model equation then reads as

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

(4)

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

(5)

with

$$\begin{pmatrix} u_{D^+} \\ u_{D^-} \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{D^+} & \rho \sigma_{D^-} \sigma_{D^+} \\ \rho \sigma_{D^-} \sigma_{D^+} & \sigma^2_{D^-} \end{pmatrix} \right),$$  

(6)

where $y_{D^+}$ and $y_{D^-}$ denote the diagnostic test values in the population of diseased and non-diseased of the study, respectively. The study-specific (indexed by $i$, $i \in 1, ..., N$) random effects $u_{D^+}$ and $u_{D^-}$ with their corresponding variances $\sigma^2_{D^+}$ and $\sigma^2_{D^-}$ and correlation parameter $\rho$ 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 $\mu_{D^+}$, $\phi_{D^+}$, $\mu_{D^-}$, and $\phi_{D^-}$ of the Weibull, log-normal, and log-logistic distribution. Additionally, we define $\lambda_{D^+}$ and $\lambda_{D^-}$ as the scale parameters of the error terms $\epsilon_{D^+}$ and $\epsilon_{D^-}$, which will be used explicitly for predicting sensitivities and specificities as presented in Section 3.3. As given in Hoyer et al.,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: $b_{D^+} = -\phi_{D^+}^{-1} \log(\mu_{D^+})$ and $\lambda_{D^+} = 1/\phi_{D^+}$. 

Finally, $\epsilon_{D^+}$ and $\epsilon_{D^-}$ are the corresponding error terms that define the distributions after the log-transformation of $y_{D^+}$ and $y_{D^-}$. These distributions are the Gumbel, normal, and logistic distribution with expected value 0 and scale parameters $\lambda_{D^+}$ and $\lambda_{D^-}$ in case the Weibull, log-normal, and log-logistic distribution are used on the original scale, respectively.
Parameter estimation of the model is based on the maximum likelihood principle while appropriately accounting for interval-censored data as, for example, presented by Kalbfleisch and Prentice. Therefore, it is possible to use every statistical software that allows for maximizing likelihood functions with bivariate random effects.

Borrowing methods from time-to-event analysis leads to various advantages as we now have a model that uses the complete information on all evaluated thresholds given in the single studies. Moreover, the bivariate time-to-event model enables us to predict sensitivities and specificities at various thresholds by simply using survival functions. Additionally, we are not restricted to studies that report on the same numbers and values of functions. The central idea of our proposed extension of the model presented in Section 3.1 is therefore to use the bivariate time-to-event model, but we now allow for more flexible distributions of the underlying diagnostic test values and also for comparison of different models via AIC or BIC.

We use the parametrization for the GF family as proposed by Prentice and Cox. The probability density function for the population of diseased \( D^+ \) is

\[
\begin{align*}
    f_{GF}(y_{D^+}) &= \frac{\partial_{D^+} \exp(-b_{D^+} m_{1D^+} / \lambda_{D^+}) y_{D^+}^{(\partial_{D^+} / \lambda_{D^+}) m_{1D^+} + (m_{1D^+} / m_{2D^+}) m_{1D^+}}}{y_{D^+} \partial_{D^+} B(m_{1D^+}, m_{2D^+}) \left[1 + (m_{1D^+} / m_{2D^+}) (\exp(-b_{D^+} y_{D^+}) y_{D^+}^{(\partial_{D^+} / \lambda_{D^+}) m_{1D^+} + m_{1D^+} \right],
\end{align*}
\]

Equations (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 \( \partial_{D^+} \) are defined as follows:

\[
\begin{align*}
    m_{1D^+} &= 2 \left(q_{D^+}^2 + 2 p_{D^+} + q_{D^+} (q_{D^+}^2 + 2 p_{D^+})^{1/2}\right)^{-1}, \quad (8) \\
    m_{2D^+} &= 2 \left(q_{D^+}^2 + 2 p_{D^+} - q_{D^+} (q_{D^+}^2 + 2 p_{D^+})^{1/2}\right)^{-1}, \quad (9)
\end{align*}
\]

and

\[
\partial_{D^+} = (m_{1D^+}^{-1} + m_{2D^+}^{-1})^{1/2} = (q_{D^+}^2 + 2 p_{D^+})^{1/2}, \quad (10)
\]

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 four parameters \( p, q, b, \lambda \) and is therefore very flexible. Thereby, \( b \) and \( \lambda \) represent the location and scale parameter, respectively, and \( m_1 \) and \( m_2 \) as transformations of \( p \) and \( q \), denote the, potentially non-integer, degrees of freedom. Moreover, by restricting \( p \) and \( q \) 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.
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. Focusing 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 Hoyer et al\textsuperscript{11} for arbitrary diagnostic test values $y$:

- for the Weibull distribution,
  \[
  \text{Sensitivity} = \exp\left( - \left( y \exp(-b_D^+) \right)^{1/\lambda_D^+} \right); \tag{11}
  \]
  where $\Phi$ is the cdf of the normal distribution; and

- for the log-normal distribution,
  \[
  \text{Sensitivity} = 1 - \Phi\left( \frac{1}{\lambda_D^+} (\log(y) - b_D^+) \right), \tag{12}
  \]

- for the log-logistic distribution,
  \[
  \text{Sensitivity} = 1 / \left( 1 + \exp\left( \frac{-b_D^+}{\lambda_D^+} \right) \right)^{-1}; \tag{13}
  \]

For the remaining members of the GF family, we obtain\textsuperscript{19} for arbitrary diagnostic test values $y$:

- for the generalized log-logistic distribution,
  \[
  \text{Sensitivity} = 1 - F_{\text{Beta}}(x_{\text{GLL}}, a_{\text{GLL}}, b_{\text{GLL}}), \tag{14}
  \]
  where $F_{\text{Beta}}$ is the cumulative distribution function (cdf) of the beta distribution with

  \[
  x_{\text{GLL}} = \frac{1/p_D^+}{1/p_D^+ + 1/p_D^+ \exp\left( -2/2(\log(y) - b_D^+) \right)}, \quad \text{and}
  \]
  \[
  a_{\text{GLL}} = b_{\text{GLL}} = 1/p_D^+; \quad \text{and}
  \]
  - for the Burr III distribution,
    \[
    \text{Sensitivity} = 1 - F_{\text{Beta}}(x_{\text{BIII}}, a_{\text{BIII}}, b_{\text{BIII}}), \tag{15}
    \]
    where $x_{\text{BIII}} = 1 + m_{2D^+} \exp(s_D^+ (\log(y) - b_{D^+} + b_{D^+}) / \lambda_D^+)$, $a_{\text{BIII}} = 1$, $b_{\text{BIII}} = m_{1D^+}$, $q_{D^+} = - (1 - p_D^+) \sqrt{2} / 2 - p_D^+$, and $m_{1D^+} = (\sqrt{\lambda_D^+} + 2p_D^+ - q_{D^+} \sqrt{\lambda_D^+} + 2p_D^+)^{-1}$;

- for the Burr XII distribution,
  \[
  \text{Sensitivity} = 1 - F_{\text{Beta}}(x_{\text{BXII}}, a_{\text{BXII}}, b_{\text{BXII}}), \tag{16}
  \]
  where $x_{\text{BXII}} = m_{1D^+} + m_{2D^+} \exp(s_D^+ (\log(y) - b_{D^+} + b_{D^+}) / \lambda_D^+)$, $a_{\text{BXII}} = m_{2D^+}$, $b_{\text{BXII}} = 1$, the parameters $q_D^+$ and $m_{2D^+}$ being defined as $q_D^+ = (1 - p_D^+) \sqrt{2} / 2 - p_D^+$ and $m_{2D^+} = (\sqrt{\lambda_D^+} + 2p_D^+ - q_{D^+} \sqrt{\lambda_D^+} + 2p_D^+)^{-1}$;

- for the generalized $F$ distribution,
  \[
  \text{Sensitivity} = 1 - F_{\text{Beta}}(x_{\text{GF}}, a_{\text{GF}}, b_{\text{GF}}), \tag{17}
  \]
  where $x_{\text{GF}} = m_{2D^+} + m_{2D^+} \exp(s_D^+ (\log(y) - b_{D^+} + b_{D^+}) / \lambda_D^+)$, $a_{\text{GF}} = m_{2D^+}$, $b_{\text{GF}} = m_{2D^+}$ with
  \[
  m_{1D^+} = 2\left(q_{D^+}^2 + 2p_D^+ + q_D^+ \left(q_D^2 + 2p_D^+\right)^{1/2}\right)^{-1} \quad \text{and}
  \]
  \[
  m_{2D^+} = 2\left(q_{D^+}^2 + 2p_D^+ - q_D^+ \left(q_D^2 + 2p_D^+\right)^{1/2}\right)^{-1}. \quad \text{and}
  \]

4 SIMULATION

To compare the model using the generalized log-logistic, Burr III, Burr XII, and GF distribution with the
previously used Weibull, log-normal, and log-logistic approach,\textsuperscript{11} we conducted a simulation study. The corresponding simulation program was written in SAS 9.4 (SAS Institute Inc., Cary, North Carolina).

4.1 Setting

For our simulation study, simulated data were drawn from the Weibull, log-normal, log-logistic, and GF distribution. That means, these four distributions served as our “true underlying distributions.” We used the estimates of \(b_D^+, b_D^-, \lambda_D^+\), and \(\lambda_D^-\) for the Weibull, log-normal, and log-logistic distribution and, 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 \(\rho\) 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.

4.2 Data generation

Combining all parameter settings led to 12 different simulation scenarios. We simulated 1000 meta-analyses for each of them. In line with Hoyer et al,\textsuperscript{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 Hoyer et al,\textsuperscript{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 Hoyer et al.\textsuperscript{11} That means, we first generated a bivariate normally distributed random effect according to Equation (6) with predefined \(\sigma^2_D^+, \sigma^2_D^-, \text{ and } \rho\). With respect to Equations (4) and (5), this random effect was added to \(b_D^-, \text{ and } b_D^+\). In combination with previously defined \(\lambda_D^-\) and \(\lambda_D^+\), 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 four-fold tables.

For generating random numbers from the GF distribution, we adapted the algorithm implemented in the \textit{flexsurv} package of the statistical software R\textsuperscript{20} to SAS. This algorithm is based on the representation of the GF family as given in Cox\textsuperscript{19} and inverse transform sampling.

| 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%         |
4.3 Estimation methods and outcomes

For each of the 1000 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 Hoyer et al 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 submodel 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 follows:

\[
AIC = 2f(\hat{\theta}) + 2k,
\]

and

\[
BIC = 2f(\hat{\theta}) + k\log(N),
\]

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 \( N \) the number of studies.

4.4 Results

In the following, we report on the results of our simulation study, which is focused 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 Supporting Information. Tables 5 and 6 show the complete results concerning numerical robustness and model selection.

4.4.1 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 (ie, 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 with 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.

4.4.2 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 with 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.

4.4.3 Number of converged runs

In terms of numerical robustness, we report on the total number of converged simulation runs (out of 1000) 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 was 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.

4.4.4 Model selection criteria

In Table 6, we report on the number that each model has been selected because of the minimal AIC or BIC. The

| True Model/ Correlation | Threshold | Sens | Spec |
|-------------------------|-----------|------|------|
| Generalized F/0.28      | 5.0       | –0.4 | 14.8 |
|                         | 5.5       | –1.0 | 1.3  |
|                         | 6.0       | –1.4 | –4.0 |
|                         | 6.5       | 0.4  | 1.9  |
|                         | 7.0       | 2.7  | 1.7  |

| True Model/ Correlation | Threshold | Sens | Spec |
|-------------------------|-----------|------|------|
| Generalized F/0.28      | 5.0       | –5.7 | 3.3  |
|                         | 5.5       | –10.9| 9.0  |
|                         | 6.0       | –19.0| 6.2  |
|                         | 6.5       | –25.2| 2.6  |
|                         | 7.0       | –18.2| 1.0  |

| True Model/ Correlation | Threshold | Sens | Spec |
|-------------------------|-----------|------|------|
| Generalized F/0.28      | 5.0       | 92.3 | 4.9  |
|                         | 5.5       | 92.4 | 82.5 |
|                         | 6.0       | 93.2 | 89.1 |
|                         | 6.5       | 92.1 | 74.2 |
|                         | 7.0       | 92.6 | 2.3  |

| True Model/ Correlation | Threshold | Sens | Spec |
|-------------------------|-----------|------|------|
| Generalized F/0.28      | 5.0       | 100  | 100  |
|                         | 5.5       | 100  | 100  |
|                         | 6.0       | 50.0 | 66.7 |
|                         | 6.5       | 0.0  | 33.3 |
|                         | 7.0       | 0.0  | 20.0 |

---

**TABLE 3** Bias (in percentage points): sensitivity (sens) and specificity (spec); bold entries indicate the smallest bias

**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)
maximum that can be achieved is therefore 1000. 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 correla-
tion 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 distributions
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.

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. Focusing on the esti-
mated sensitivities and specificities at various thresholds
as the parameters with most practical relevance, the
Correlation

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

| True Model/ Correlation | Estimated Model | Generalized log-logistic | Burr III | Burr XII | Generalized F |
|------------------------|-----------------|--------------------------|----------|----------|---------------|
| Weibull/0.00           | 999             | 1000                     | 1000     | 998      | 956           | 761 | 644 |
| Log-normal/0.00        | 1000            | 1000                     | 1000     | 998      | 956           | 761 | 644 |
| Log-logistic/0.00      | 921             | 998                      | 1000     | 960      | 997           | 1000 | 966 |
| Generalized F/0.00     | 943             | 1000                     | 1000     | 783      | 8             | 653 | 871 |
| Generalized F/0.85     | 948             | 999                      | 1000     | 748      | 7             | 709 | 877 |

At the currently recommended HbA\textsubscript{1c} threshold of
6.5,\textsuperscript{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\textsubscript{1c} is an appropriate primary
screening test for type 2 diabetes, we are more interested
in higher sensitivities that are associated with lower diag-
nostic thresholds. Moreover, even in case sensitivity and
specificity are weighted equally (which corresponds to a
choice based on the Youden index), the optimal thresh-
hold 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
(AUCs) 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 two parameters, also their AUCs are largely identical.

In the Supporting Information, we present an SAS macro that can be used to fit the generalized log-logistic, Burr III, Burr XII, and the GF model.

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. This approach allows us to use information on each and every threshold reported by the single studies included in a meta-analysis. In contrast to existing approaches that can also use this information, 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 four 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-

| Table 6 | Model selection: number of selected models (in terms of AIC and BIC) from 1000 meta-analyses; bold entries indicate the maximum number per row and criterion |
| --- | --- | --- | --- | --- | --- | --- |
| True Model/Correlation | Estimated Model | 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 | 4 | 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/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 |

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion.
normal, and log-logistic used previously. With respect to often missing individual participant data in the meta-analysis of full ROC curves, this flexibility is an advantage because a wide range of possible underlying distributions is covered. Second, we are now able to apply model selection criteria as the AIC or the BIC to choose the most appropriate model. Note that relying on likelihood ratio tests 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\textsuperscript{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 modeling.

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 with 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 four 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

![Figure 2](image-url)

**Figure 2** Estimated summary receiver operating characteristic (ROC) curves of the different time-to-event models. Light gray solid lines and dots depict the estimated sensitivities, specificities, and ROC curves of the single studies.

| Estimated Model       | AIC       | BIC       | m\textsubscript{1D}⁺ | m\textsubscript{1D}⁻ | m\textsubscript{2D}⁺ | m\textsubscript{2D}⁻ | AUC       | [95% CI]     | [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%     | [75.3%-85.4%]|
| Burr III              | 209 315   | 209 330   | 0.81                | 1.16                | -                   | -                   | 83.9%     | [82.3%-85.5%]|
| Burr XII              | 209 325   | 209 340   | -                   | -                   | 1.45                | 0.93                | 83.8%     | [81.9%-85.0%]|
| Generalized F         | 209 296   | 209 314   | 1.43                | 1.97                | 2.42                | 1.52                | 83.9%     | [82.3%-84.7%]|

Abbreviations: AIC, Akaike information criterion; AUCs, areas under the curve; BIC, Bayesian information criterion.
distributions as the normal or logistic distribution (as for example reported by Steinhauser et al9) instead of members of the GF family of distributions. However, especially, the GF distribution is very flexible (see, for example, Cox19) and can cope with a wide range of skewness, thus 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.

Highlights

What is already known:

- Methods for the meta-analysis of diagnostic studies frequently focused 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 ROC have been proposed; however, they come with several disadvantages.

What is new:

- We extend the previously used bivariate time-to-event model for interval-censored data for the meta-analysis of full ROC curves by referring to the flexible family of generalized F (GF) distributions, which is governed by four parameters and includes all of the previously used distributions as special cases. As a consequence, model selection criteria as the AIC or the BIC to choose the most appropriate model are applicable.

Potential impact for RSM readers outside the authors’ field

- In this article, we showed how methods from survival analysis can be used for the meta-analysis of full ROC curves based on a flexible set of distributions. The proposed model can be used in various medical applications to evaluate diagnostic tests.

CONFLICT OF INTEREST

The author reported no conflict of interest.

DATA AVAILABILITY STATEMENT

The complete data set presented in Section 2 was also used in our previous publication where it is given as Supporting Information.11 Additionally, the data set is available on request from the authors.

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. https://doi.org/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. https://doi.org/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. Martinez-Cambor P. Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Stat Method Med Res*. 2017;26:5-20. https://doi.org/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. https://doi.org/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. *Res Synth 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. https://doi.org/10.1002/dmr.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. https://doi.org/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. https://doi.org/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.

ORCID

Annika Hoyer https://orcid.org/0000-0002-0241-9951
17. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. Hoboken, NJ: John Wiley & Sons; 2011.
18. Prentice RL. Discrimination among some parametric models. *Biometrika*. 1975;62:607-614.
19. Cox C. The generalized $F$ distribution: an umbrella for parametric survival analysis. *Stat Med*. 2008;27:4301-4312.
20. Jackson CH. Flexsurv: a platform for parametric survival modeling in R. *J Stat Softw*. 2016;70:1-33.

How to cite this article: Hoyer A, Kuss O. Meta-analysis of full ROC curves with flexible parametric distributions of diagnostic test values. *Res Syn Meth*. 2020;11:301–313. [https://doi.org/10.1002/jrsm.1395](https://doi.org/10.1002/jrsm.1395)