The Annals of Applied Statistics
2019, Vol. 13, No. 1, 548–572
https://doi.org/10.1214/18-AOAS1207

MODELING BIOMARKER RATIOS WITH GAMMA DISTRIBUTED COMPONENTS[1]

BY MORITZ BERGER*, MICHAEL WAGNER*,† AND MATTHIAS SCHMID*,†

University of Bonn and German Center for Neurodegenerative Diseases*†

We propose a regression model termed "extended GB2 model", which is designed to analyze ratios of biomarkers in epidemiological and medical research. Typical examples of biomarker ratios are given by the LDL/HDL cholesterol ratio in cardiovascular research and the amyloid-β 42/40 ratio in dementia research. Unlike regression modeling with a log-transformed response, which is often used to describe ratio outcomes in observational studies, the extended GB2 model directly links the expectation of the untransformed biomarker ratio to a set of covariates. This strategy allows for a simple interpretation of the predictor-response relationships in terms of multiplicative increases/decreases of the expected outcome, similar to Poisson and Cox regression. In the theoretical part of the paper, we derive the log-likelihood of the proposed model, analyze its properties, and provide details on confidence intervals and hypothesis testing. We will also present the results of a simulation study demonstrating the robustness of the proposed modeling approach against model misspecification. The usefulness of the method is demonstrated by two applications on the aforementioned LDL/HDL cholesterol and amyloid-β 42/40 ratios. For this, we analyze data from a cohort study on kidney disease and from a large observational database on neurodegenerative diseases.

1. Introduction. This paper presents a regression model for the ratio of two positively correlated biomarkers U and V, which are assumed to follow a joint bivariate distribution with gamma distributed components. The model relates the expectation of the ratio U/V to a set of explanatory variables $X = (X_1, \dots, X_p)^\top$, allowing for model building and inference in the same way as in generalized linear models (GLMs).

Outcome variables of the type U/V are frequently encountered in observational studies. Important examples, which will also be considered here, are the LDL/HDL cholesterol ratio in cardiovascular research [Natarajan et al. (2003)] and the amyloid-β 42/40 ratio in dementia research [Koyama et al. (2012)]. The LDL/HDL cholesterol ratio is defined as the ratio of the low-density lipoprotein

Received January 2018; revised August 2018.

[1] Supported by grants from the German Federal Ministry of Education and Research (Kompetenznetz Demenzen, Grant number 01GI0420, and GCKD Study, Grant number 01ER0804) and the German Research Foundation, Grant SCHM 2966/1-2.

Key words and phrases. Biomarker ratio, gamma distribution, generalized beta distribution of the second kind, generalized linear model, regression.

(LDL) and the high-density lipoprotein (HDL) concentrations in plasma or serum. Both concentrations can be described by right-skewed random variables with positive correlation. Since high values of LDL/HDL are a strong predictor of cardiovascular events [Natarajan et al. (2003), Millan et al. (2009)], extensive epidemiological research is conducted to study the effect of lifestyle factors, nutritional components, dietary patterns, and other nutrition-related parameters (such as body mass index) on LDL/HDL cholesterol levels [Müller et al. (2003), Weggemans and Trautwein (2003), Sacks et al. (2006), Sundram, Karupaiah and Hayes (2007), Shamai et al. (2011)]. In dementia research, the amyloid-β 42/40 ratio is defined as the ratio of the amyloid-β 42 protein level and the amyloid-β 40 isoform concentration in cerebrospinal fluid or plasma. Again, both concentrations can be described by right-skewed random variables with positive correlation. In recent years, the amyloid-β 42/40 ratio has been identified as a diagnostic and prognostic factor for the progression of Alzheimer's disease [AD, Koyama et al. (2012), Lewczuk et al. (2015)]. Importantly, decreases in the amyloid-β 42/40 ratio are considered to be an early phenomenon in AD progression that is often evident in patients long before the first clinical symptoms of AD. It is therefore of considerable interest to model the effects of dementia-related factors such as age and education on the amyloid-β 42/40 ratio, and to investigate the characteristics and progression of AD in its early stages.

Another important example is the CD4/CD8 ratio in HIV research, which measures the ratio of T helper cells to cytotoxic T cells in the human immune system. A low CD4/CD8 ratio indicates ongoing immune activation and affects the risk of non-AIDS morbidity and mortality. Recently, a persistent CD4/CD8 ratio < 1 has been found to be associated with several risk factors, like a low CD4 T-cell nadir and a shorter duration of viral suppression [Caby et al. (2016)].

A common approach to model biomarker ratios in observational studies is to apply a logarithmic transformation to the variable U/V and to fit a regression model with log-transformed outcome $\log(U/V)$. This approach can be considered as a special case of the more general theory on log-ratio transformations used in compositional data analysis [Aitchison (1986), Wang and Zhao (2017)]. A well-known property of log-ratio analysis is that inference is not possible for the conditional expectation $\mathrm{E}[U/V|X]$ but only for the expectation on the log-transformed scale $\mathrm{E}[\log(U/V)|X]$. This affects the interpretability of the predictor-response relationships and the conclusions drawn from the results of associated hypothesis tests.

An alternative approach, which forms the basis of the method developed here, is to model right-skewed positive data by gamma distributed random variables. This strategy is justified by the result that analyzing log-normal data assuming a gamma distribution is often more efficient than analyzing gamma data assuming log-normality [see Wiens (1999) and Firth (1988), who compared formulas for the asymptotic relative efficiencies of the two approaches]. However, while gamma regression for the positive random variables U and V is widely used in practice, only few models for the *ratio* of two gamma distributed random variables U and V

exist. If independence between U and V can be assumed, the ratio U/V follows a generalized beta distribution of the second kind [“GB2 distribution”, Kleiber and Kotz (2003)]. For the latter distribution, Tulupyev et al. (2013) proposed a GLM-type regression model, which, however, was originally not intended to analyze ratio outcomes but to adjust a time-to-event model for a specific sampling pattern. As a consequence, the model by Tulupyev et al. (2013) makes strong assumptions about the parameter space, restricting one of the two shape parameters to the value 2. A more general model for the GB2 distribution was proposed by Yee (2015) as part of the VGAM framework. Regarding the ratio of *correlated* gamma distributed variables, which is of particular interest in biomarker research [Long et al. (2016)], no regression modeling strategy for $\mathrm{E}[U/V|X]$ exists (to the best of our knowledge). In fact, although there are several well-established results on the probability density function (p.d.f.) and the moments of U/V, these have not been incorporated in a GLM-type regression framework yet. In part, this may be due to the fact that the p.d.f. and the moments of U/V involve mathematical expressions that are not available as analytic formulas [Lee, Holland and Flueck (1979), Tubbs (1986)], making iterative procedures for maximum likelihood estimation infeasible.

To address these issues, we propose the *extended GB2 model*, which relates the expectation of the untransformed ratio $\mathrm{E}[U/V]$ directly to the covariates X. The log-likelihood function of the model is derived from a bivariate gamma distribution for (U, V) [“Kibble–Wicksell distribution”, Kibble (1941)], which allows for possibly different means and variances of the components U and V. The extended GB2 model also accounts for correlations between U and V, with one of the model parameters being directly interpretable in terms of the Pearson correlation coefficient ρ. For $\rho = 0$, the extended GB2 model reduces to the standard GB2 model for independent variables U and V (hence the name *extended* GB2 model).

A major advantage of the proposed modeling strategy is that it is possible to derive an analytic formula for the p.d.f. of $(U/V)|X$. This allows for maximum likelihood estimation and hypothesis testing in the same way as in GLMs. The estimates of the model parameters have a simple interpretation in terms of multiplicative increases/decreases of the expected ratio $\mathrm{E}[U/V|X]$.

The rest of the paper is organized as follows: Section 2 introduces notation and definitions, and provides details on the derivation and the interpretation of the extended GB2 model. In addition, we will show how to construct GLM-type confidence intervals and hypothesis tests. Section 3 presents two applications from observational research dealing with the analysis of the aforementioned LDL/HDL cholesterol and amyloid-β 42/40 ratios. Modeling of the LDL/HDL cholesterol ratio will be based on data collected for the German Chronic Kidney Disease (GDKD) Study, which is one of the largest cohort studies worldwide to analyze CKD patients without dialysis treatment [Titze et al. (2015)]. To model the amyloid-β 42/40 ratio, we will use data collected for the Dementia Competence

Network (DCN), which maintains a large observational database on neurodegenerative diseases [Kornhuber et al. (2009)]. In Section 4 we will summarize the results of two simulation studies that were carried out to investigate the finite-sample properties of the extended GB2 model. In these studies, which will be presented in more detail in the supplemental article of the paper [Berger, Wagner and Schmid (2019)], we will compare the extended GB2 model to the Gaussian model with log-transformed outcome variable and analyze its robustness with respect to model misspecification. Section 5 summarizes the main findings of the paper. The proposed method has been implemented in the R add-on package **eGB2** [Berger and Schmid (2019)], which is part of the Supplementary Material of the paper.

2. Methods.

2.1. *Notation and distributional results.* Let U and V be two gamma distributed random variables with probability density functions

$$f_U(u) = \frac{\lambda_u^\alpha}{\Gamma(\alpha)} u^{\alpha-1} \exp(-\lambda_u u), \tag{1}$$

$$f_V(v) = \frac{\lambda_v^\alpha}{\Gamma(\alpha)} v^{\alpha-1} \exp(-\lambda_v v), \tag{2}$$

where $\alpha > 0$ denotes a common shape parameter and $\lambda_u, \lambda_v > 0$ are the rate parameters of f_U and f_V, respectively. The assumption of a common shape parameter for f_U and f_V ensures that the density functions in (1) and (2) share the same basic form. The means and variances of U, V are given by α/λ_u, α/λ_v, and α/λ_u^2, α/λ_v^2, respectively.

To derive a model for the ratio U/V, we assume that the pair (U, V) follows a bivariate gamma distribution with probability density function

$$\begin{aligned} f_{U,V}(u, v) = {} & \frac{(\lambda_u\lambda_v)^\alpha}{(1-\rho)\Gamma(\alpha)} \left(\frac{uv}{\rho\lambda_u\lambda_v}\right)^{\frac{\alpha-1}{2}} \\ & \times \exp\left(-\frac{\lambda_u u + \lambda_v v}{1-\rho}\right) \mathrm{I}_{\alpha-1}\left(\frac{2\sqrt{\rho\lambda_u\lambda_v uv}}{1-\rho}\right), \end{aligned} \tag{3}$$

where $0 < \rho < 1$ and $\mathrm{I}_{\alpha-1}(\cdot)$ is the modified Bessel function of the first kind of order $\alpha - 1$. The p.d.f. defined in (3) was first introduced by Kibble [Kibble (1941)] and is known as "Kibble's bivariate gamma distribution" or "Kibble–Wicksell distribution" [Balakrishnan and Lai (2009)]. It can be shown that the additional parameter ρ equals the Pearson correlation coefficient of U and V.

The p.d.f. of the ratio $R := U/V$ is derived as follows:

PROPOSITION 1. *Let the joint distribution of* (U, V) *be defined by the probability density function in* (3). *Then the p.d.f. of the random variable* $R := U/V$ *is* *given by*

$$f_R(r;\alpha,\rho,\theta) = \frac{\Gamma(2\alpha)}{\Gamma^2(\alpha)}\theta^{-\alpha}(1-\rho)^{\alpha} \times \frac{(\frac{1-\theta}{\theta}\frac{r}{1+r}+1)(\frac{r^{\alpha-1}}{(1+r)^{2\alpha}})}{((\frac{1-\theta}{\theta}\frac{r}{1+r}+1)^2 - 4\rho\frac{r}{\theta(1+r)^2})^{\alpha+0.5}}, \tag{4}$$

where $\theta := \lambda_v/\lambda_u = E[U]/E[V]$ *denotes the ratio of the two rate parameters.*

PROOF. See Appendix A. □

By Proposition 1, the p.d.f. in (4) defines a distribution for the ratio of two correlated gamma distributed random variables with possibly different means and variances. For $\rho = 0$ the p.d.f. in (4) reduces to

$$f_R(r;\alpha,\theta) = \frac{\Gamma(2\alpha)}{\Gamma^2(\alpha)}\theta^{-\alpha}r^{\alpha-1}\left(\frac{r}{\theta}+1\right)^{-2\alpha}, \tag{5}$$

which corresponds to the p.d.f. of the *generalized beta distribution of the second kind*. This p.d.f. is obtained when U and V are independent [Kleiber and Kotz (2003)]. Illustrations of the p.d.f. of $R = U/V$ are shown in Figure 1.

The expectation of R can be expressed as follows:

PROPOSITION 2. *Under the assumptions of Proposition 1, the expectation of the random variable* $R = U/V$ *is given by*

$$\begin{aligned} E[U/V] &= \frac{E[U]}{E[V]} \cdot \frac{\Gamma(\alpha+1)\Gamma(\alpha-1)}{\Gamma^2(\alpha)} {}_2F_1(-1,1;\alpha;\rho) \\ &= \theta\frac{\Gamma(\alpha+1)\Gamma(\alpha-1)}{\Gamma^2(\alpha)} {}_2F_1(-1,1;\alpha;\rho) \\ &= \theta C(\alpha,\rho), \qquad \alpha > 1, \end{aligned} \tag{6}$$

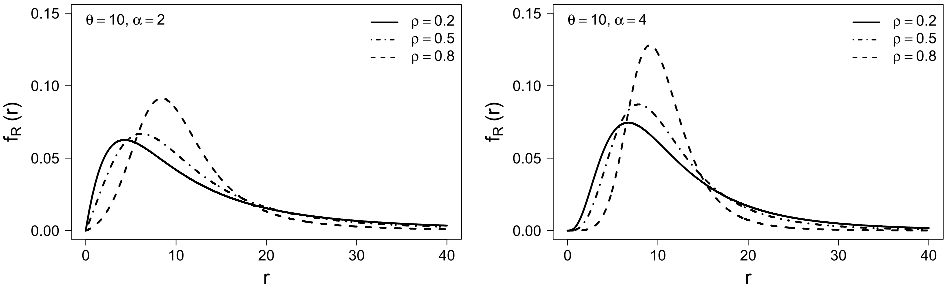


FIG. 1. *Examples of the p.d.f. of the ratio* $R = U/V$ *derived in Proposition 1. The curves visualize the distribution of R for parameters* $\alpha \in \{2, 4\}$, $\rho \in \{0.2, 0.5, 0.8\}$ *and* $\theta = 10$.

where $C(\alpha, \rho) := \Gamma(\alpha+1)\Gamma(\alpha-1)/\Gamma^2(\alpha)\,{}_2F_1(-1, 1; \alpha; \rho)$ *and* ${}_2F_1(\cdot)$ *is the generalized hypergeometric function.*

PROOF. Proposition 2 follows from the application of Corollary 1 in Candan and Orguner (2013), who derived a general expression for the moment function of U/V. □

Proposition 2 implies that the expectation of $R = U/V$ increases with the value of the correlation coefficient ρ (see also Figure 1). Since the existence of E[R] is not guaranteed if $\alpha \leq 1$, we will assume $\alpha > 1$ when deriving the extended GB2 model in the next section.

2.2. *Definition and interpretation of the extended GB*2 *model.* By Proposition 2, the mean ratio E[U/V] can be written as the product of the ratio of means $\theta = \mathrm{E}[U]/\mathrm{E}[V]$ and a factor $C(\alpha, \rho)$ that is a function of the parameters α and ρ. Based on these considerations, we relate θ to $X_1, \dots, X_p$ by the model equation

$$\log(\theta|X) = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p, \tag{7}$$

where $\gamma := (\gamma_0, \gamma_1, \dots, \gamma_p)^\top$ is a set of real-valued coefficients. The logarithmic transformation in (7) projects the positive random variable $\theta|X$ to the set of real numbers, ensuring that there are no restrictions on the coefficients $(\gamma_0, \gamma_1, \dots, \gamma_p)^\top$. In the following, we will refer to the model defined in (7) as *extended GB2 model*. Note that it was implicitly assumed in (7) that categorical covariates are represented by dummy-coded variables or an equivalent coding.

PROPOSITION 3. *Under the assumption that the shape parameter* α *and the correlation coefficient* ρ *do not depend on* X, *the extended GB*2 *model is equivalent to the following model for the mean ratio* $E[U/V|X]$:

$$\log\bigl(E[U/V|X]\bigr) = \tilde{\gamma}_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p, \tag{8}$$

where $\tilde{\gamma}_0 := \gamma_0 + \log(C(\alpha, \rho))$.

PROOF. Proposition 3 follows from equation (7) and from the proportionality between θ and E[U/V] stated in Proposition 2. □

Proposition 3 shows that the predictor-response relationships defined by $\gamma_1, \dots, \gamma_p$ are the same for the mean ratio E[$U/V|X$] and the ratio of means $\theta|X = \mathrm{E}[U|X]/\mathrm{E}[V|X]$, provided that α and ρ can be treated as nuisance parameters. More specifically, since (7) can be re-written as

$$\theta|X = \exp(\gamma_0) \cdot \exp(\gamma_1 X_1) \cdot \cdots \cdot \exp(\gamma_p X_p), \tag{9}$$

the expressions $\exp(\gamma_1), \dots, \exp(\gamma_p)$ have a simple interpretation in terms of multiplicative increases/decreases of the expected ratio E[$U/V|X$]. For example, if $\gamma_k > 0$, $k \in \{1, \dots, p\}$, increasing X_k by one unit implies that E[$U/V|X$] is increased by the factor $\exp(\gamma_k)$.

2.3. *Maximum likelihood estimation*, *inference*, *and point predictions*. Estimates of the parameters $\gamma_0, \gamma_1, \dots, \gamma_p$ are obtained by maximizing the log-likelihood of the extended GB2 model. Let $(r_i, x_{1i}, \dots, x_{pi})^\top$, $i = 1, \dots, n$, be a set of independent realizations of $(R, X_1, \dots, X_p)^\top$ and define $x_i := (1, x_{1i}, \dots, x_{pi})^\top$. Then, according to Proposition 1 and equation (7), the log-likelihood of the extended GB2 model is given by

$$\begin{aligned} & l(\gamma, \alpha, \rho; r_1, \dots, r_n, x_1, \dots, x_n) \\ & \quad = \sum_{i=1}^{n} \biggl\{ \log(\Gamma(2\alpha)) - 2\log(\Gamma(\alpha)) - \alpha \cdot x_i^\top \gamma + \alpha \log(1-\rho) \\ & \qquad + \log\biggl((\exp(-x_i^\top \gamma) - 1)\frac{r_i}{r_i + 1} + 1 \biggr) + (\alpha - 1)\log(r_i) \\ & \qquad - 2\alpha \log(1 + r_i) - (\alpha + 0.5)\biggl(\log\biggl(\biggl((\exp(-x_i^\top \gamma) - 1)\frac{r_i}{r_i + 1} + 1 \biggr)^2 \\ & \qquad - 4\rho \exp(-x_i^\top \gamma)\frac{r_i}{(r_i + 1)^2} \biggr) \biggr) \biggr\}. \end{aligned} \tag{10}$$

By standard maximum likelihood arguments, consistent estimators of the model parameters are defined by

$$(\hat{\gamma}^\top, \hat{\alpha}, \hat{\rho})^\top := \underset{\gamma, \alpha, \rho}{\operatorname{argmax}}\, l(\gamma, \alpha, \rho; r_1, \dots, r_n, x_1, \dots, x_n). \tag{11}$$

Details on the numerical optimization of (10) are provided below.

Because of the asymptotic normality of the maximum likelihood estimators, statistical tests of the hypotheses "H_0: $\gamma_k = 0$ vs. H_1: $\gamma_k \neq 0$", $k = 1, \dots, p$, are obtained by plugging the estimates in the observed information matrix $J(\gamma, \alpha, \rho) := -\partial^2 l(\gamma, \alpha, \rho; r_1, \dots, r_n, x_1, \dots, x_n)/\partial\gamma\gamma^\top$ and by calculating the test statistics

$$Z_k = \frac{\hat{\gamma}_k}{\sqrt{J_{kk}^{-1}(\hat{\gamma}, \hat{\alpha}, \hat{\rho})}}, \qquad k \in \{1, \dots, p\}, \tag{12}$$

where J_{kk}^{-1} denotes the kth diagonal element of J^{-1}. For a given type I error level α_I, the null hypothesis "H_0: $\gamma_k = 0$" is rejected if $|Z_k| > z_{1-\alpha_\mathrm{I}/2}$, where $z_{1-\alpha_\mathrm{I}/2}$ is the $(1 - \alpha_\mathrm{I}/2)$-quantile of the standard normal distribution. More general linear hypotheses of rank $\tilde{r} \geq 1$ (e.g., associated with several covariates or a factor variable with more than two levels) can be investigated by using standard log-likelihood ratio test statistics (denoted by LR) that asymptotically follow chi-squared distributions with $\tilde{r}$ degrees of freedom. P-values are defined by $2 \cdot \min\{\mathrm{P}(Z_k \leq Z_{k,\mathrm{obs}}), \mathrm{P}(Z_k \geq Z_{k,\mathrm{obs}})\}$, and $\mathrm{P}(LR \geq LR_\mathrm{obs})$ under the respective null hypotheses, where $Z_{k,\mathrm{obs}}$ and LR_obs denote the observed values of Z_k and LR, respectively.

Asymptotic $(1-\alpha_{\mathrm{I}})\%$ Wald confidence intervals for the parameters γ_k are defined by $\hat{\gamma}_k \pm z_{1-\alpha_{\mathrm{I}}/2}\sqrt{J_{kk}^{-1}(\hat{\gamma}, \hat{\alpha}, \hat{\rho})}$. Alternatively, it is possible to construct $(1-\alpha_{\mathrm{I}})\%$ profile likelihood confidence intervals, which are given by the sets of parameters γ_k^0 for which the log-likelihood ratio test statistic (comparing the full model and the model under the null hypothesis $\gamma_k = \gamma_k^0$) does not exceed the $(1-\alpha_{\mathrm{I}})$-quantile of the chi-squared distribution with one degree of freedom. Comparisons of Wald and profile likelihood confidence intervals will be presented in the two applications in Section 3.

The maximum likelihood estimates also allow for computing point predictions for the expected ratio $\mathrm{E}[U/V|X]$. According to Proposition 3, these are given by

$$\widehat{\mathrm{E}}[U/V|x_i] = \exp(\hat{\gamma}_1 x_{1i}) \cdot \dots \cdot \exp(\hat{\gamma}_p x_{pi}) \cdot \exp(\hat{\gamma}_0) \cdot C(\hat{\alpha}, \hat{\rho}). \tag{13}$$

Maximization of the log-likelihood function (10) over γ can be carried out using the R package **eGB2** [Berger and Schmid (2019)], which also allows for fitting the simple GB2 model with correlation coefficient $\rho = 0$. The optimization algorithm used by **eGB2** is based on the implementation of the "Broyden, Fletcher, Goldfarb, and Shanno" (BFGS) algorithm in the R function `optim()`. Profile likelihood confidence intervals can be computed via the function `profileCI()`. Alternatively, **eGB2** offers to use a gradient boosting algorithm with component-wise linear base-learners [Hofner et al. (2014)] for maximization of the log-likelihood. Details on gradient boosting are presented in Section 1 of the supplemental article.

2.4. *Quasi-likelihood and relationship to gamma regression.* According to Proposition 2 and Corollary 1 in Candan and Orguner (2013), the extended GB2 model is characterized by the mean-variance relationship

$$\mathrm{E}[R] = \theta C(\alpha, \rho), \tag{14}$$

$$\begin{aligned} \mathrm{Var}[R] &= \mathrm{E}[R^2] - (\mathrm{E}[R])^2 \\ &= \theta^2 C_2(\alpha, \rho) - \theta^2 C(\alpha, \rho)^2 \\ &= \theta^2 (C_2(\alpha, \rho) - C(\alpha, \rho)^2) \\ &= \mathrm{E}[R]^2 C_3(\alpha, \rho), \end{aligned} \tag{15}$$

where $C_2(\alpha, \rho) := \Gamma(\alpha+2)\Gamma(\alpha-2)/\Gamma^2(\alpha)_2F_1(-2, 2; \alpha; \rho)$ and $C_3(\alpha, \rho) := C_2(\alpha, \rho)/C(\alpha, \rho)^2 - 1$. This relationship can be used to fit a "quasi extended GB2 model", which is helpful when the fully parametric extended GB2 model does not fit the data well and the underlying assumptions on the p.d.f. of the ratio U/V given in (4) need to be relaxed.

Up to a multiplicative factor, the mean-variance relationship in (14) and (16) is the same as the one defined by an ordinary gamma regression model with outcome variable $R^* \sim \text{Gamma}(\alpha^*, \lambda^*)$, link function

$$\log(\mathrm{E}[R^*|X]) = \log(\alpha^*) + \gamma_0^* + \gamma_1^* X_1 + \cdots + \gamma_p^* X_p =: \log(\alpha^*) + \log(\theta^*), \tag{16}$$

and mean-variance relationship

$$\mathrm{E}[R^*] = \theta^* \alpha^*, \tag{17}$$

$$\begin{aligned} \mathrm{Var}[R^*] &= (\theta^*)^2 \alpha^* \\ &= \mathrm{E}[R^*]^2 C^*(\alpha^*), \end{aligned} \tag{18}$$

where $\lambda^* = 1/\theta^*$ and $C^*(\alpha^*) := 1/\alpha^*$. It follows that the quasi-likelihood functions of the two models are equivalent, and that software designed for fitting quasi gamma models can be used for fitting quasi extended GB2 models as well. In particular, the quasi-likelihood function defined by the mean-variance relationship in (17) and (19) yields the same maximum quasi-likelihood estimates as the one defined by the quadratic mean-variance relationship $\mu := \mathrm{E}[R^*]/\alpha^*$ and $\mathrm{V}(\mu) := \mathrm{Var}[R^*]/\alpha^* = \mu^2$ [McCullagh and Nelder (1989), pp. 325 ff.]. The quasi extended eGB2 models considered in this paper are therefore based on the quadratic mean-variance relationship $\mathrm{V}(\mu) = \mu^2$. For details on quasi-likelihood estimation, see in particular McCullagh and Nelder (1989), Chapter 9.

REMARK. The mean-variance relationship in (14) and (16) also allows for fitting a quasi extended GB2 model to longitudinal data via generalized estimation equations (GEE). This kind of analysis additionally requires to specify an appropriate working correlation matrix representing within-subject dependencies. An example using longitudinal data collected for the GCKD Study is presented in Appendix B.

3. Applications.

3.1. *German chronic kidney disease study.* To illustrate the application of the extended GB2 model, we analyzed the LDL/HDL cholesterol ratios collected in the German Chronic Kidney Disease (GCKD) Study [Titze et al. (2015)]. The GCKD Study is an ongoing multi-center cohort study that enrolled 5217 patients with either stage III chronic kidney disease or overt albuminuria/proteinuria between March 2010 and March 2012. Data collection comprised measurements on clinical variables (e.g., renal function), lifestyle factors (e.g., smoking behavior and alcohol consumption), and laboratory measurements obtained from blood and urine samples (in particular, serum LDL and HDL cholesterol concentrations). For details on the inclusion/exclusion criteria and the design of the study, we refer to Eckardt et al. (2012) and Titze et al. (2015).

One of the aims of the GCKD Study is to investigate the risk of cardiovascular events in patients with renal disease. Since LDL and HDL cholesterol concentrations are associated with a number of cardiovascular diseases [CVD, Mendis, Puska and Norrving (2011)] it is of high interest to relate the LDL/HDL cholesterol ratios of CVD-free study participants to explanatory variables such as age, smoking behavior, body mass index, and renal function.

Here we focus on the LDL/HDL cholesterol ratios of the subgroup of GCKD Study participants that neither suffered from coronary heart disease at baseline nor experienced one or more previous strokes. This resulted in a data set with $n = 3669$ patients. There were missing values in 4.3% of the patients (see the description in Table 1). The largest part of these missing values (94 out of 157 patients) occurred in the HbA1c levels, which could not be determined in all patients due to insufficient sample volumes or possible problems with the analytes during laboratory analysis. A test on random missingness ["MAR+" assumption, Potthoff et al. (2006)] did not lead us to suspect any violations of the missing-at-random assumption ($p = 0.9790$, for details on the test see Section 2 of the supplemental article). After the exclusion of patients with missing values in any of the analyzed variables, we obtained an analysis data set with $n = 3512$ participants. Figure 2 shows

TABLE 1

Description and summary statistics of the variables used for the analysis of the baseline LDL/HDL cholesterol ratios in the GCKD Study ($Q1$ = *first quartile*, $Q3$ = *third quartile*). *High levels of urinary albumin and/or low levels of the estimated glomerular filtration rate indicate a decreased renal function. HbA1c levels* $> 6.5\%$ *are an indicator of diabetes. Out of the* $n = 3669$ *patients that neither suffered from coronary heart disease nor experienced one or more previous strokes*, 157 *patients* (4.3%) *had missing values in at least one of the analyzed variables. For details on the collection of the data, see Eckardt et al.* (2012) *and Titze et al.* (2015)

	Summary statistics						
Variable	**min**	**Q1**	**median**	**Q3**	**max**	**mean**	**sd**
LDL/HDL cholesterol ratio	0.20	1.75	2.35	3.09	13.22	2.51	1.06
Age (years)	18	50	61	69	76	58.24	12.67
Body mass index (kg/m^2)	15.50	25.30	28.40	32.80	69.70	29.46	6.08
Urinary albumin (g/l)	0.002	0.006	0.041	0.300	17.44	0.329	0.833
Est. glomerular filtration rate (ml/min per 1.73 m^2)	8	36	45	55	151	48.34	17.64
Gender	male:		1964 (56.0%)		female:		1548 (44.0%)
Heavy alcohol consumption	no:		2841 (80.9%)		yes:		671 (19.1%)
HbA1c level > 6.5%	no:		2553 (72.7%)		yes:		959 (27.3%)
Smoking	no:		1373 (39.1%)		former:		1559 (44.4%)
	yes:		580 (16.5%)				
Physical activity	0:		583 (16.6%)		1–2:		894 (25.4%)
(times per week)	3–5:		1014 (28.9%)		>5:		1021 (29.1%)

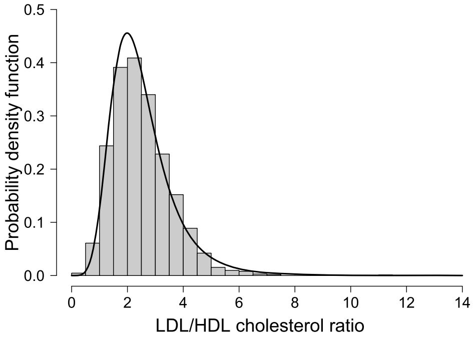


FIG. 2. *Distribution of the baseline LDL/HDL cholesterol ratios measured in the GCKD study* ($n = 3512$). *The black line refers to the p.d.f. of the ratio* $R = U/V$ *defined in* (4). *It was estimated by fitting a covariate-free extended GB*2 *model to the GCKD baseline data.*

the unconditional distribution of the LDL/HDL ratios. The unconditional Pearson correlation coefficient between the LDL and HDL cholesterol levels was 0.161. Summary statistics of the variables used in the analysis are presented in Table 1.

Table 2 presents the results obtained from fitting the extended GB2 model to the GCKD baseline data using the BFGS algorithm. The coefficient estimates and p-values confirm various established results on the associations between clinical variables, lifestyle factors, and cardiovascular disease. For example, body mass index, which is a major CVD risk factor, was estimated to increase the expected LDL/HDL cholesterol ratio by the factor exp(0.0115) = 1.0116 (i.e., by 1.16%) per kg/m^2. A similar result was obtained for physical activity, which lowered the expectation of the LDL/HDL cholesterol ratio ($\hat{\gamma} \leq -0.0312$). Table 2 also confirms the positive association between alcohol consumption and HDL cholesterol [Linn et al. (1993), $\hat{\gamma} = -0.0886$, $p < 0.0001$]. The effects of the renal parameters urinary albumin ($\hat{\gamma} = 0.0191$, $p = 0.0326$) and estimated glomerular filtration rate ($\hat{\gamma} = -0.0007$, $p = 0.0976$) confirm the previously established association between chronic kidney disease and CVD [Gansevoort et al. (2013)].

To investigate the goodness-of-fit of the extended GB2 model, we computed the quantile residuals of the fitted model [Dunn and Smyth (1996)] and compared them to the respective quantiles of a standard normal distribution (Figure 3). Although the distribution of the residuals shows slight deviations from normality, Figure 3 does not indicate any substantial problems with the fit of the extended GB2 model. Also, the profile likelihood confidence intervals presented in Table 2 were similar to the respective Wald intervals, indicating that the quadratic approximation of the extended GB2 log-likelihood worked well.

3.2. *Cohort study of the German dementia competence network.* In a second application, we analyzed data from the multi-center observational cohort study conducted by the German Dementia Competence Network [DCN, Kornhuber et

TABLE 2

*Analysis of the LDL/HDL cholesterol ratios in the German Chronic Kidney Disease Study. Coefficient estimates, p-values and confidence intervals were obtained by fitting an extended GB*2 *model to the GCKD baseline data using the BFGS algorithm* ($\hat{\gamma}$ = *coefficient estimate, CI* = *confidence interval*). *Reference categories of categorical covariates are indicated by dots. The maximum likelihood estimates of* α *and* ρ *were* 4.766 *and* 0.658, *respectively*

			95% CI		
	$\hat{\gamma}$	**exp(**$\hat{\gamma}$**)**	**Wald**	**Profile**	p**-value**
Age (years)	−0.0011	0.9989	[0.9977; 1.0001]	[0.9977; 1.0001]	0.0676
Body mass index (kg/m^2)	0.0115	1.0116	[1.0093; 1.0139]	[1.0093; 1.0139]	<0.0001
Urinary albumin (g/l)	0.0191	1.0193	[1.0016; 1.0374]	[1.0018; 1.0376]	0.0326
Est. glomerular filtration rate (ml/min per 1.73 m^2)	−0.0007	0.9993	[0.9985; 1.0001]	[0.9985; 1.0001]	0.0976
Gender (male)	.				
Gender (female)	−0.1764	0.8383	[0.8142; 0.8631]	[0.8148; 0.8637]	<0.0001
Heavy alcohol consumption (no)	.				
Heavy alcohol consumption (yes)	−0.0886	0.9153	[0.8827; 0.9490]	[0.8825; 0.9488]	<0.0001
HbA1c level ≤6.5% (no)	.				
HbA1c level >6.5% (yes)	0.0127	1.0128	[0.9823; 1.0442]	[0.9828; 1.0417]	0.4140
Smoking (non-smoker)	.				
Smoking (former)	−0.0483	0.9528	[0.9242; 0.9824]	[0.9242; 0.9824]	0.0004
Smoking (yes)	0.0242	1.0245	[0.9841; 1.0665]	[0.9842; 1.0529]	
Physical activity (0)	.				
Physical activity (1–2)	−0.0312	0.9692	[0.9287; 1.0115]	[0.9468; 1.0102]	0.2358
Physical activity (3–5)	−0.0408	0.9600	[0.9208; 1.0009]	[0.9210; 1.0014]	
Physical activity (>5)	−0.0397	0.9611	[0.9219; 1.0019]	[0.9450; 1.0026]	

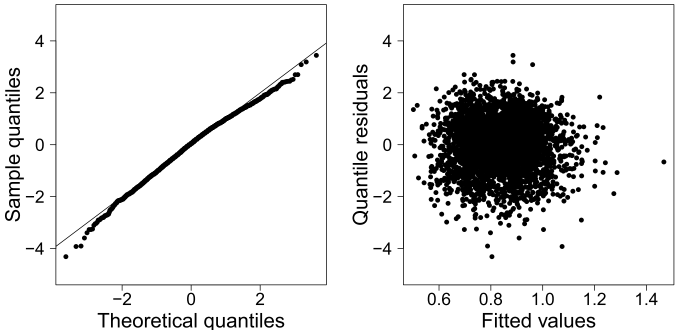


FIG. 3. *Analysis of the baseline LDL/HDL cholesterol ratios in the German Chronic Kidney Disease Study. The left panel shows a plot of the quantile residuals obtained from the extended GB*2 *model against the quantiles of a standard normal distribution. The right panel shows a plot of the quantile residuals against the fitted values of the extended GB*2 *model.*

al. (2009)]. The study included patients older than 50 years who sought evaluation at one of the participating university memory clinics. Dementia-related diagnoses were either mild cognitive impairment (MCI), Alzheimer's disease (AD), or other dementia. All diagnoses were made using clinical and neuropsychological assessments.

The aims of the study were to establish the diagnostic and prognostic power of clinical, laboratory, and imaging methods. Cerebrospinal fluid (CSF) samples were collected, and a variety of laboratory parameters (in particular, amyloid-β 42 and 40 protein concentrations) were measured. Baseline data collection took place between 2003 and 2007. For details on the assessment procedures, we refer to Kornhuber et al. (2009).

A major challenge in the diagnosis and prognosis of AD is the decades-long period between disease onset and the first clinical symptoms of AD [Sperling, Karlawish and Johnson (2013)]. This problem is further aggravated by the fact that not all patients passing through the MCI stage will eventually suffer from underlying AD pathology [Kornhuber et al. (2009)]. It has therefore been suggested to use biomarkers such as amyloid-β 42 protein concentrations for early AD diagnosis and prediction, and to relate them to AD risk factors such as age and level of education. Since the amyloid-β 42/40 ratio is considered to be a stronger predictor of AD progression than amyloid-β 42 concentrations alone [Wiltfang et al. (2007), Koyama et al. (2012)], it is of high interest to relate amyloid-β 42/40 measurements to dementia-related risk factors in MCI patients.

Here we focus on the amyloid-β 42/40 ratios of the DCN Study participants at baseline, which were available in 380 of the 1095 patients diagnosed with MCI. The reason for this reduction in sample size was the incomplete number of CSF biosamples, which were not collected from all patients due to either logistic reasons or lack of consent to the invasive procedure of lumbar puncture. Of note, the biomarker sampling rate of the DCN cohort is comparable to that of other observational MCI memory clinic cohorts in the field of AD, for example, the ADNI Study [Kornhuber et al. (2009)]. Out of the 380 patients, seven patients were excluded from analysis because they did not meet the inclusion criteria (age ≤ 50 years). Out of the remaining 373 patients, 37 patients had missing values in one or more of the analyzed covariates; 35 of these patients had a missing value in the ApoE $\epsilon 4$ covariate (defined in more detail below), which could not be measured for the same reasons as those stated above. Again, testing the MAR+ assumption [Potthoff et al. (2006)] did not lead us to suspect nonrandom missingness ($p = 0.1442$, for details on the test see Section 2 of the supplemental article). Exclusion of the 37 patients with missing values resulted in an analysis data set with $n = 336$ patients. Figure 4 shows the unconditional distribution of the amyloid-β 42/40 ratios. The unconditional Pearson correlation coefficient between the amyloid-β 42 and the amyloid-β 40 concentrations was 0.422.

The following covariates were considered for inclusion in the extended GB2 model: (i) gender, (ii) age in years, (iii) educational level (measured by the number

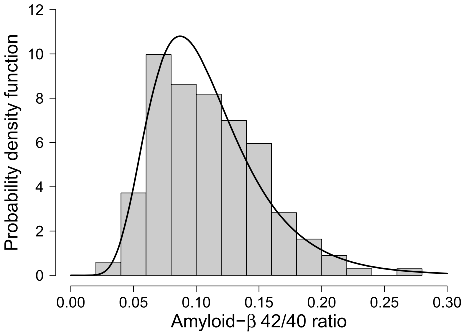


FIG. 4. *Distribution of the amyloid-β 42/40 ratios measured in patients with MCI (DCN study, $n = 336$). The black line refers to the p.d.f. of the ratio $R = U/V$ defined in (4). It was estimated by fitting a covariate-free extended GB2 model to the DCN study data collected at baseline.*

of years of education), and (iv) a binary variable indicating whether a patient was a carrier of the apolipoprotein E $\epsilon 4$ (ApoE $\epsilon 4$) allele, which is a strong genetic predictor of AD. Summary statistics of the variables used in the analysis are presented in Table 3.

Table 4 presents the results obtained from fitting the extended GB2 model to the DCN baseline data using the BFGS algorithm. According to the p-values given in the last column of Table 4, the AD risk factors age and ApoE $\epsilon 4$ showed strong evidence for an effect on the amyloid-β 42/40 ratios of the study participants ($p = 0.0002$ and $p < 0.0001$, respectively). Each year of age was estimated to reduce the expected amyloid-β 42/40 ratio by the factor $\exp(-0.0093) = 0.9907$, corresponding to a yearly decrease of approximately 1%. Expected amyloid-β 42/40 ratios of ApoE $\epsilon 4$ carriers were reduced by an estimated 18% compared to patients not carrying the allele ($\exp(-0.2040) = 0.8154$). These results confirm

TABLE 3

Description and summary statistics of the variables used for the analysis of the baseline amyloid-β 42/40 ratios in the DCN cohort study ($Q1$ = first quartile, $Q3$ = third quartile). All numbers refer to a subset of patients diagnosed with MCI ($n = 336$). For details on the collection of the data, see Kornhuber et al. (2009)

	Summary statistics						
Variable	**min**	**Q1**	**median**	**Q3**	**max**	**mean**	**sd**
Amyloid-β 42/40 ratio	0.03	0.07	0.10	0.14	0.26	0.11	0.04
Age (years)	51	60	66	72	89	66.45	8.10
Education (years)	2	11	12	13	19	12.17	2.95
Gender	male:		197 (58.7%)		female:	139 (41.3%)	
ApoE $\epsilon 4$	no:		185 (55.1%)		yes:	151 (44.9%)	

TABLE 4

Analysis of the amyloid-β 42/40 ratios in the cohort study of the German Dementia Competence Network. Coefficient estimates, p-values, and confidence intervals were obtained by fitting an extended GB2 model to the DCN baseline data using the BFGS algorithm ($\hat{\gamma}$ = coefficient estimate, CI = confidence interval). Reference categories of categorical covariates are indicated by dots. The maximum likelihood estimates of α and ρ were 9.607 and 0.379, respectively

			95% CI		
	$\hat{\gamma}$	$\exp(\hat{\gamma})$	Wald	Profile	p-value
Age (years)	−0.0093	0.9907	[0.9858; 0.9956]	[0.9858; 0.9957]	0.0002
Education (years)	0.0001	1.0001	[0.9864; 1.0140]	[0.9864; 1.0140]	0.9882
Gender (male)	.				
Gender (female)	−0.0596	0.9421	[0.8673; 1.0234]	[0.9116; 1.0230]	0.1577
ApoE ϵ4 (no)	.				
ApoE ϵ4 (yes)	−0.2040	0.8155	[0.7524; 0.8838]	[0.7525; 0.8797]	<0.0001

the negative associations between the two AD risk factors and amyloid pathology [e.g., Jack et al. (2015)]. In contrast to age and ApoE ϵ4, there was no evidence for an effect of educational level or gender on the amyloid-β 42/40 ratio.

The quantile residuals presented in Figure 5 indicate a very good fit of the extended GB2 model. Again, the profile likelihood confidence intervals were similar to the respective Wald intervals (see Table 4).

3.3. *Comparison of models.* The evaluation of the quantile residuals in Sections 3.1 and 3.2 suggests that the extended GB2 model fitted the data well in both applications. For comparison, we investigated the performance of the extended GB2 model in terms of prediction accuracy and compared it to the performance

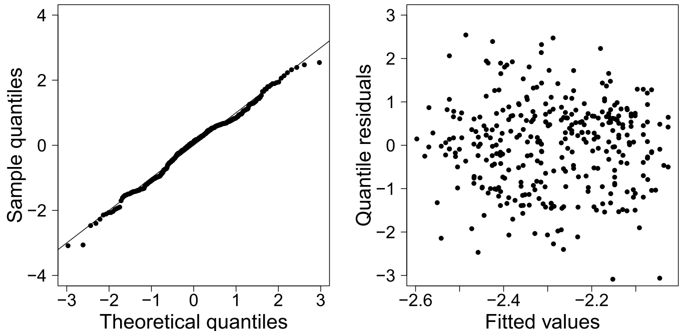
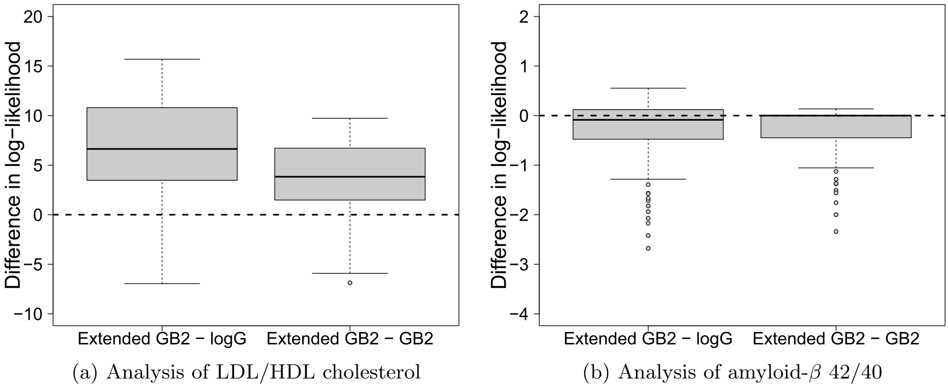


FIG. 5. *Analysis of the baseline amyloid-β 42/40 ratios in the cohort study of the German Dementia Competence Network. The left panel shows a plot of the quantile residuals obtained from the extended GB2 model against the quantiles of a standard normal distribution. The right panel shows a plot of the quantile residuals against the fitted values of the extended GB2 model.*

of competing models. To do so, we first generated 100 random subsamples without replacement from both data sets and fitted regression models to all subsamples. Up to rounding errors, each subsample comprised two thirds of the original GCKD/DCN baseline data. The following models were considered: (i) the extended GB2 model, (ii) a Gaussian model with log-transformed outcome variable (abbreviated by "logG" in the following), and (iii) a simple GB2 model with correlation parameter $\rho = 0$. In the next step, we evaluated the model fits by computing predictive log-likelihood values from the remaining patients not used for model fitting (100 subsamples comprising one third of the original GCKD/DCN data each, up to rounding errors). According to the results presented in Figure 6, the extended GB2 model performed best when analyzing the LDL/HDL cholesterol ratios in the GCKD Study. In the DCN Study, the differences between the three models were less distinct. Wilcoxon signed-rank tests suggested that all median differences in Figure 6 were significantly different from zero ($p < 0.001$ in all four tests). Since the p-values of these tests depend on the number of subsamples (with no guideline on how to choose an "optimal" number of subsamples being available), we additionally calculated Akaike's information criterion (AIC) from the models. For the GCKD Study, we obtained AIC values of 9615.455 (extended GB2), 9658.704 (logG), and 9640.651 (simple GB2), resulting in AIC differences of 43.249 (logG − extended GB2) and 25.196 (simple GB2 − extended GB2). For the DCN Study, we obtained AIC values of −1234.727 (extended GB2), −1236.634 (logG), and −1236.411 (simple GB2), resulting in AIC differences of 1.907 (extended GB2 − logG) and 1.684 (extended GB2 − simple GB2). Hence, according to the rules of thumb provided in Burnham and Anderson (2002), Section 2.6, the AIC differences obtained from the GCKD study suggest "essentially no empirical support"



(a) Analysis of LDL/HDL cholesterol

(b) Analysis of amyloid-β 42/40

FIG. 6. *Analysis of the baseline data of the GCKD Study (left panel) and the DCN Study (right panel). All models were fitted to* 100 *subsamples without replacement of sizes* $n_{\text{train}} = 2341$ *(GCKD data) and* $n_{\text{train}} = 224$ *(DCN data) each. Predictive log-likelihood values were computed from the* 100 *subsamples not used for model fitting* ($n_{\text{test}} = 1171$ *and* $n_{\text{test}} = 112$ *for the GCKD and DCN data, respectively).*

of the logG and the simple GB2 models compared to the extended GB2 model (AIC differences > 10), whereas in the DCN study the extended GB2 model had "substantial empirical support" compared to the logG and the simple GB2 models (AIC differences < 2).

4. Simulations.

To further analyze the properties of the extended GB2 model, we conducted two simulation studies. The aims of the studies were (i) to analyze the model fit and the power of the hypothesis tests, and (ii) to compare the extended GB2 model to the logG model. In *Simulation Study 1*, the data were generated according to the p.d.f. of the extended GB2 model. Several scenarios with varying correlation coefficients ρ were considered. In *Simulation Study 2*, the data were generated according to the p.d.f. of a log-normal distribution. The following models were fitted to the data: (i) the extended GB2 model, (ii) the logG model, (iii) the quasi extended GB2 model, and (iv) a simple GB2 model with $\rho = 0$.

The results of the two studies can briefly be summarized as follows: In Simulation Study 1, the finite-sample bias of the maximum likelihood estimates of the extended GB2 model was througout small, regardless of the value of ρ. Also, the estimated power of the hypothesis tests was higher in the extended GB2 model than in the simple GB2 model. The differences between the two models increased with the value of ρ, confirming that the efficiency of the estimators increases when the correlation between the ratio components U and V is taken into account. The performance of the quasi extended GB2 model was comparable for small values of ρ, but deteriorated already for moderate values of ρ. The rejection rates of the logG model were almost identical to those of the simple GB2 model. Regarding the goodness of model fit, the extended GB2 model performed throughout better than the logG model in Simulation Study 1. This result, which was expected due to the extended GB2 model being the true data-generating model in Simulation Study 1, was particularly evident when ρ was large. In Simulation Study 2, the extended GB2 model resulted in similar model fits as the logG model, despite the latter model being the true data-generating model. For details on the simulation studies, see Section 3 of the supplemental article [Berger, Wagner and Schmid (2019)]. All the results can be reproduced using the supplementary files [Berger (2018)].

5. Summary and conclusion.

The extended GB2 model is a regression modeling approach that is applicable whenever the outcome of interest is a ratio of two variables with right-skewed distributions and positive correlation. In the applications presented in Section 3, we used the model to analyze outcomes defined by the ratio of a biomarker of interest (LDL cholesterol/amyloid-β 42 protein) and a positively associated "reference" marker (HDL cholesterol/amyloid-β 40 protein).

According to the theoretical and empirical results presented in Sections 2 to 4, the main advantages of the extended GB2 model are:

Interpretability. The proposed modeling approach is derived from a well-defined theoretical model for the ratio of two correlated gamma distributed variables. Hence it reflects the true underlying composition of the biomarker ratio. In addition, the extended GB2 model allows for a simple interpretation of the parameter estimates in terms of multiplicative increases/decreases of the expected ratio outcome. This interpretation is analogous to Poisson and Cox regression and is thus familiar to many researchers.

Applicability. Parameter estimates of the extended GB2 model are easily obtained by maximum likelihood estimation. Furthermore, the proposed modeling approach results in confidence intervals, hypothesis tests, and point predictions of the conditional expectation of the ratio outcome. As demonstrated in Section 3, the fit of the model can be assessed by computing and plotting quantile residuals.

Increase in Efficiency. Compared to the simple GB2 model that ignores the correlation between the ratio components, the extended GB2 model increases both the goodness of the model fit and the power of the associated hypothesis tests.

Robustness. The simulation study presented in Section 4 and the Supplementary Material suggests that the extended GB2 model is fairly robust against model misspecification. This is in line with earlier findings by Firth (1988) and Wiens (1999). In addition, as described in Section 2.4, it is possible to use a quasi-likelihood modeling approach when the assumptions of the fully parametric extended GB2 model are not satisfied.

Extensibility. There are numerous extensions and additional modeling options for the extended GB2 model. For example, it is straightforward to extend the model by interaction terms and nonlinear predictor effects [e.g., modeled via P-splines, Eilers and Marx (1996)]. Furthermore, it is possible to increase the flexiblity of the model by relating the parameters α and ρ to the covariates. This strategy would embed the extended GB2 model in the *generalized additive models for location, scale, and shape* (GAMLSS) framework developed by Rigby and Stasinopoulos (2005). In higher-dimensional settings, variable selection could be carried out using AIC- or BIC-based methods, or by modifying the gradient boosting algorithm described in Section 1 of the supplemental article using early stopping [e.g., Hofner et al. (2014)]. Furthermore, it is straightforward to fit a GEE version of the quasi extended GB2 model in order to analyze clustered and/or longitudinal data (see the analysis of the GCKD follow-up data in Appendix B). Embedding the extended GB2 model in the framework of generalized mixed-effects models appears to be possible but will require more complex optimization methods and is subject to further research.

Finally we emphasize that we do *not* consider the extended GB2 model to be a “generally better” modeling option than Gaussian log-ratio modeling. While the purpose of this paper is to introduce the extended GB2 model and to illustrate its application to real-world data, considerably more work is required to conduct an in-depth comparison of the various modeling approaches for biomarker ratios and other ratio outcomes.

APPENDIX A: PROOF OF PROPOSITION 1

To prove Proposition 1, we first consider the random variable $W := U/(U+V)$ with realization $w \in (0,1)$ and probability density function $f_W(w)$. Under the assumption that (U,V) follows the bivariate gamma distribution defined in Equation (3) of the paper, it can be shown that

$$\begin{aligned} f_W(w;\alpha,\rho,\theta) &= \frac{\Gamma(2\alpha)}{\Gamma^2(\alpha)}\theta^{-\alpha}(1-\rho)^{\alpha} \\ &\quad \times \frac{(\frac{1-\theta}{\theta}w+1)(w(1-w))^{\alpha-1}}{((\frac{1-\theta}{\theta}w+1)^2-\frac{4\rho}{\theta}w(1-w))^{\alpha+0.5}}, \end{aligned} \tag{19}$$

see Nadarajah and Kotz (2007) and Weinhold et al. (2016) for details.

To derive the p.d.f. of the random variable $R = U/V$ with realization $r = u/v$, we define the transformation function $\phi(w) := w/(1-w) = u/v = r$. Since ϕ is strictly monotonically increasing in w, it follows that the p.d.f. of R can be written as

$$f_R(r;\alpha,\rho,\theta) = f_W(\phi^{-1}(r))\left|\frac{\partial}{\partial r}\phi^{-1}(r)\right|, \tag{20}$$

where

$$\phi^{-1}(r) = \frac{r}{1+r} \tag{21}$$

and

$$\left|\frac{\partial}{\partial r}\phi^{-1}(r)\right| = \frac{1}{(1+r)^2} \tag{22}$$

are the inverse of ϕ and its derivative, respectively.

Combining (19) to (22) yields the p.d.f. of the ratio R:

$$\begin{aligned} &f_R(r;\alpha,\rho,\theta) \\ &\quad = f_W\left(\frac{r}{1+r}\right)\frac{1}{(1+r)^2} \\ &\quad = \frac{\Gamma(2\alpha)}{\Gamma^2(\alpha)}\theta^{-\alpha}(1-\rho)^{\alpha}\frac{1}{(1+r)^2} \\ &\qquad \cdot \frac{(\frac{1-\theta}{\theta}\cdot\frac{r}{1+r}+1)(\frac{r}{1+r}(1-\frac{r}{1+r}))^{\alpha-1}}{((\frac{1-\theta}{\theta}\cdot\frac{r}{1+r}+1)^2-\frac{4\rho}{\theta}\cdot\frac{r}{1+r}(1-\frac{r}{1+r}))^{\alpha+0.5}} \\ &\quad = \frac{\Gamma(2\alpha)}{\Gamma^2(\alpha)}\theta^{-\alpha}(1-\rho)^{\alpha}\frac{1}{(1+r)^2} \end{aligned} \tag{23}$$

$$\cdot \frac{(\frac{1-\theta}{\theta} \cdot \frac{r}{1+r} + 1)(\frac{r}{(1+r)^2})^{\alpha-1}}{((\frac{1-\theta}{\theta} \cdot \frac{r}{1+r} + 1)^2 - \frac{4\rho}{\theta} \cdot \frac{r}{(1+r)^2})^{\alpha+0.5}}$$

$$= \frac{\Gamma(2\alpha)}{\Gamma^2(\alpha)} \theta^{-\alpha}(1-\rho)^\alpha \frac{((\frac{1-\theta}{\theta} \cdot \frac{r}{1+r} + 1)(\frac{r^{\alpha-1}}{(1+r)^{2\alpha}})}{((\frac{1-\theta}{\theta} \cdot \frac{r}{1+r} + 1)^2 - 4\rho \frac{r}{\theta(1+r)^2})^{\alpha+0.5}}.$$

APPENDIX B: GEE ANALYSIS OF THE GCKD STUDY DATA

To illustrate the application of the extended GB2 model to longitudinal data, we analyzed the LDL/HDL cholesterol ratios using the baseline measurements and the measurements of the first two follow-up visits of the GCKD Study. Out of the $n = 3512$ participants in our analysis data set at baseline (Section 3.1), 3062 participants had at least one follow-up examination and 2073 participants had two follow-up examinations. Reductions in sample size were either due to deaths (121 patients between baseline and follow-up 1, and 122 patients between follow-up 1 and follow-up 2) or due to drop-outs. Exclusion of patients with missing values (following the same rationale as in Section 3.1) resulted in an analysis data set including 2813 patients with at least one follow-up visit and 1641 patients with two follow-up visits. Application of the test by Diggle (1989) did not lead us to suspect informative drop-out ($p = 0.3323$, for details on the test see Section 2 of the supplemental article).

For the GEE analysis we used the variables reported in Table 1. At the time of the analysis (June 2018) the lifestyle characteristics alcohol consumption, smoking, and physical activity were only available at baseline and were therefore treated as time-independent covariates.

Table 5 shows the results obtained from fitting the GEE version of the quasi extended GB2 model with quadratic mean-variance relationship to the GCKD baseline and follow-up data (cf. Section 2.4). The model contained separate intercept terms for baseline and the two follow-up time points. The working correlation matrix was assumed to be unstructured [argument `corstr` of the `geeglm` function in R package **geepack**, Hojsgaard, Halekoh and Yan (2016)]. It is seen that the coefficient estimates and the standard errors largely confirm the results obtained from the analysis of the GCKD baseline data (cf. Table 2).

The estimated working correlation matrix was

$$\begin{pmatrix} 1.0000 & 0.6360 & 0.5777 \\ 0.6360 & 1.0000 & 0.6570 \\ 0.5777 & 0.6570 & 1.0000 \end{pmatrix}, \tag{24}$$

revealing positive associations between the baseline measurements and the measurements at the two follow-up examinations.

TABLE 5

Longitudinal analysis of the LDL/HDL cholesterol ratios in the German Chronic Kidney Disease Study. Coefficient estimates and standard errors were obtained by fitting the GEE version of the quasi extended GB2 model to the GCKD baseline and follow-up data ($\hat{\gamma}$ = coefficient estimate, se = standard error). Reference categories of categorical covariates are indicated by dots. Standard errors refer to the robust estimates in the R add-on package **geepack**

	$\hat{\gamma}$	$\mathrm{se}(\hat{\gamma})$	$\exp(\hat{\gamma})$
Baseline	0.8491	0.0492	2.3375
Follow-up 1	0.7798	0.0492	2.1810
Follow-up 2	0.8080	0.0493	2.2434
Age (years)	−0.0015	0.0005	0.9985
Body mass index (kg/m^2)	0.0097	0.0010	1.0097
Urinary albumin (g/l)	0.0233	0.0072	1.0236
Est. glomerular filtration rate (ml/min per 1.73 m^2)	−0.0011	0.0004	0.9989
Gender (male)	.		
Gender (female)	−0.1662	0.0136	0.8469
Heavy alcohol consumption (no)	.		
Heavy alcohol consumption (yes)	−0.0843	0.0175	0.9192
HbA1c level ≤ 6.5% (no)	.		
HbA1c level > 6.5% (yes)	0.0044	0.0089	1.0044
Smoking (non-smoker)	.		
Smoking (former)	−0.0319	0.0142	0.9686
Smoking (yes)	0.0438	0.0195	1.0448
Physical activity (0)	.		
Physical activity (1–2)	0.0101	0.0117	1.0102
Physical activity (3–5)	0.0046	0.0112	1.0046
Physical activity (>5)	0.0154	0.0126	1.0155

Acknowledgements. We thank the GCKD Study Investigators for providing data of the GCKD Study for illustrative purposes. The GCKD study was further supported by a grant from the KfH Foundation for Preventive Medicine (http://www.kfhstiftung-praeventivmedizin.de/content/stiftung).

The authors thank the two anonymous reviewers and the Editor (Leonhard Held) for their thorough reading and helpful suggestions for improving the paper.

SUPPLEMENTARY MATERIAL

Supplemental Article: eGB2_supplements (DOI: 10.1214/18-AOAS1207SUPPA; .pdf). Online appendix containing details on (1) gradient boosting, (2) tests for nonrandom missingness and drop-out, and (3) simulations.

R add-on package: eGB2 (DOI: 10.1214/18-AOAS1207SUPPB; .zip). R package implementing the proposed methodology.

Supplemental Files: R Code of Section 4 (DOI: 10.1214/18-AOAS1207SUPPC; .zip). Files containing the R code to recompute the simulation studies and to reproduce the results presented in Section 4 and the supplemental article.

REFERENCES

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. CRC Press, London. MR0865647

BALAKRISHNAN, N. and LAI, C.-D. (2009). *Continuous Bivariate Distributions*, 2nd ed. Springer, Dordrecht. MR2840643

BERGER, M. (2018). Supplemental files to "Modeling Biomarker Ratios with Gamma Distributed Components." Available at https://imbie.meb.uni-bonn.de/~berger/RCode_Simulations.zip.

BERGER, M. and SCHMID, M. (2019). eGB2: Fitting (Extended) GB2 Models R package version 1.0.1. DOI:10.1214/18-AOAS1207SUPPB.

BERGER, M., WAGNER, M. and SCHMID, M. (2019). Supplement to "Modeling Biomarker Ratios with Gamma Distributed Components." DOI:10.1214/18-AOAS1207SUPPA.

BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model Selection and Multimodel Inference*: *A Practical Information-Theoretic Approach*, 2nd ed. Springer, New York. MR1919620

CABY, F., GUIHOT, A., LAMBERT-NICLOT, S., GUIGUET, M., BOUTOLLEAU, D., AGHER, R., VALANTIN, M.-A., TUBIANA, R., CALVEZ, V., MARCELIN, A.-G., CARCELAIN, G., AUTRAN, B., COSTAGLIOLA, D. and KATLAMA, C. (2016). Determinants of a low CD4/CD8 ratio in HIV-1-infected individuals despite long-term viral suppression. *Clin. Infect. Dis.* **62** 1297–1303.

CANDAN, Ç. and ORGUNER, U. (2013). The moment function for the ratio of correlated generalized gamma variables. *Statist. Probab. Lett.* **83** 2353–2356. MR3093825

DIGGLE, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics* **45** 1255–1258.

DUNN, P. K. and SMYTH, G. K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist.* **5** 236–244.

ECKARDT, K. U., BÄRTHLEIN, B., BAID-AGRAWAL, S., BECK, A., BUSCH, M., EITNER, F., EKICI, A. B., FLOEGE, J., GEFELLER, O., HALLER, H., HILGE, R., HILGERS, K. F., KIELSTEIN, J. T., KRANE, V., KÖTTGEN, A., KRONENBERG, F., OEFNER, P., PROKOSCH, H. U., REIS, A., SCHMID, M., SCHAEFFNER, E., SCHULTHEISS, U. T., SEUCHTER, S. A., SITTER, T., SOMMERER, C., WALZ, G., WANNER, C., WOLF, G., ZEIER, M. and TITZE, S. (2012). The German Chronic Kidney Disease (GCKD) study: Design and methods. *Nephrol. Dial. Transplant.* **27** 1454–1460.

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** 89–121. MR1435485

FIRTH, D. (1988). Multiplicative errors: Log-normal or gamma? *J. Roy. Statist. Soc. Ser. B* **50** 266–268. MR0964179

GANSEVOORT, R. T., CORREA-ROTTER, R., HEMMELGARN, B. R., JAFAR, T. H., HEERSPINK, H. J., MANN, J. F., MATSUSHITA, K. and WEN, C. P. (2013). Chronic kidney disease and cardiovascular risk: Epidemiology, mechanisms, and prevention. *Lancet* **27** 339–352.

HOFNER, B., MAYR, A., ROBINZONOV, N. and SCHMID, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. *Comput. Statist.* **29** 3–35. MR3260108

HOJSGAARD, S., HALEKOH, U. and YAN, J. (2016). geepack: Generalized Estimating Equation Package R package version 1.2-1. Available at https://cran.r-project.org/web/packages/geepack.

JACK, C. R., WISTE, H. J., WEIGAND, S. D., KNOPMAN, D. S., VEMURI, P., MIELKE, M. M., LOWE, V., SENJEM, M. L., GUNTER, J. L., MACHULDA, M. M., GREGG, B. E., PANKRATZ, V. S., ROCCA, W. A. and PETERSEN, R. C. (2015). Age, sex, and APOE $\varepsilon 4$ effects on memory, brain structure, and β-amyloid across the adult life span. *JAMA Neurol.* **72** 511–519.

KIBBLE, W. F. (1941). A two-variate gamma type distribution. *Sankhyā* **5** 137–150. MR0007218

KLEIBER, C. and KOTZ, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley Interscience, Hoboken, NJ. MR1994050

KORNHUBER, J., SCHMIDTKE, K., FRÖLICH, L., PERNECZKY, R., WOLF, S., HAMPEL, H., JESSEN, F., HEUSER, I., PETERS, O., WEIH, M., JAHN, H., LUCKHAUS, C., HÜLL, M., GERTZ, H. J., SCHRÖDER, J., PANTEL, J., RIENHOFF, O., SEUCHTER, S. A., RÜTHER, E., HENN, F., MAIER, W. and WILTFANG, J. (2009). Early and differential diagnosis of dementia and mild cognitive impairment: Design and cohort baseline characteristics of the German Dementia Competence Network. *Dement. Geriatr. Cogn. Disord.* **27** 404–417.

KOYAMA, A., OKEREKE, O. I., YANG, T., BLACKER, D., SELKOE, D. J. and GRODSTEIN, F. (2012). Plasma amyloid-beta as a predictor of dementia and cognitive decline—a systematic review and meta-analysis. *Archives of Neurology* **69** 824–831.

LEE, R. Y., HOLLAND, B. S. and FLUECK, J. A. (1979). Distribution of a ratio of correlated gamma random variables. *SIAM J. Appl. Math.* **36** 304–320. MR0524504

LEWCZUK, P., LELENTAL, N., SPITZER, P., MALER, J. M. and KORNHUBER, J. (2015). Amyloid-β 42/40 cerebrospinal fluid concentration ratio in the diagnostics of Alzheimer's disease: Validation of two novel assays. *J. Alzheimer's Dis.* **43** 183–191.

LINN, S., CARROLL, M., JOHNSON, C., FULWOOD, R., KALSBEEK, W. and BRIEFEL, R. (1993). High-density lipoprotein cholesterol and alcohol consumption in US white and black adults: Data from NHANES II. *Am. J. Publ. Health* **83** 811–816.

LONG, Q., ZHANG, X., ZHAO, Y., JOHNSON, B. A. and BOSTICK, R. M. (2016). Modeling clinical outcome using multiple correlated functional biomarkers: A Bayesian approach. *Stat. Methods Med. Res.* **25** 520–537. MR3489650

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. CRC Press, London. MR3223057

MENDIS, S., PUSKA, P. and NORRVING, B., eds. (2011). *Global Atlas on Cardiovascular Disease Prevention and Control*. World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization, Geneva.

MILLAN, J., PINTO, X., MUNOZ, A., ZUNIGA, M., RUBIES-PRAT, J., PALLARDO, L. F., MASANA, L., MANGAS, A., HERNANDEZ-MIJARES, A., GONZALEZ-SANTOS, P., ASCASO, J. F. and PEDRO-BOTET, J. (2009). Lipoprotein ratios: Physiological significance and clinical usefulness in cardiovascular prevention. *Vasc. Health Risk Manag.* **5** 757–765.

MÜLLER, H., LINDMAN, A. S., BRANTSAETER, A. L. and PEDERSEN, J. I. (2003). The serum LDL/HDL cholesterol ratio is influenced more favorably by exchanging saturated with unsaturated fat than by reducing saturated fat in the diet of women. *J. Nutr.* **133** 78–83.

NADARAJAH, S. and KOTZ, S. (2007). Jensen's bivariate gamma distribution: Ratios of components. *J. Stat. Comput. Simul.* **77** 349–358. MR2345738

NATARAJAN, S., GLICK, H., CRIQUI, M., HOROWITZ, D., LIPSITZ, S. R. and KINOSIAN, B. (2003). Cholesterol measures to identify and treat individuals at risk for coronary heart disease. *Am. J. Prev. Med.* **25** 50–57.

POTTHOFF, R. F., TUDOR, G. E., PIEPER, K. S. and HASSELBLAD, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Stat. Methods Med. Res.* **15** 213–234. MR2227446

RIGBY, R. A. and STASINOPOULOS, D. M. (2005). Generalized additive models for location, scale and shape. *J. Roy. Statist. Soc. Ser. C* **54** 507–554. MR2137253

SACKS, F. M., LICHTENSTEIN, A., VAN HORN, L., HARRIS, W., KRIS-ETHERTON, P. and WINSTON, M. (2006). Soy protein, isoflavones, and cardiovascular health: An American Heart Association Science Advisory for professionals from the Nutrition Committee. *Circulation* **113** 1034–1044.

SHAMAI, L., LURIX, E., SHEN, M., NOVARO, G. M., SZOMSTEIN, S., ROSENTHAL, R., HERNANDEZ, A. V. and ASHER, C. R. (2011). Association of body mass index and lipid profiles: Evaluation of a broad spectrum of body mass index patients including the morbidly obese. *Obes. Surg.* **21** 42–47.

SPERLING, R. A., KARLAWISH, J. and JOHNSON, K. A. (2013). Preclinical Alzheimer disease—the challenges ahead. *Nat. Rev. Neurol.* **9** 54–58.

SUNDRAM, K., KARUPAIAH, T. and HAYES, K. C. (2007). Stearic acid-rich interesterified fat and trans-rich fat raise the LDL/HDL ratio and plasma glucose relative to palm olein in humans. *Nutr. Metab.* **4** 3.

TITZE, S., SCHMID, M., KÖTTGEN, A., BUSCH, M., FLOEGE, J., WANNER, C., KRONENBERG, F. and ECKARDT, K. U. (2015). Disease burden and risk profile in referred patients with moderate chronic kidney disease: Composition of the German Chronic Kidney Disease (GCKD) cohort. *Nephrol. Dial. Transplant.* **30** 441–451.

TUBBS, J. D. (1986). Moments for a ratio of correlated gamma variates. *Comm. Statist. Theory Methods* **15** 251–259. MR0828616

TULUPYEV, A., SUVOROVA, A., SOUSA, J. and ZELTERMAN, D. (2013). Beta prime regression with application to risky behavior frequency screening. *Stat. Med.* **32** 4044–4056. MR3102433

WANG, T. and ZHAO, H. (2017). Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann. Appl. Stat.* **11** 771–791. MR3693546

WEGGEMANS, R. M. and TRAUTWEIN, E. A. (2003). Relation between soy-associated isoflavones and LDL and HDL cholesterol concentrations in humans: A meta-analysis. *Eur. J. Clin. Nutr.* **57** 940–946.

WEINHOLD, L., WAHL, S., PECHLIVANIS, S., HOFFMANN, P. and SCHMID, M. (2016). A statistical model for the analysis of beta values in DNA methylation studies. *BMC Bioinform.* **17** 480.

WIENS, B. L. (1999). When log-normal and gamma models give different results: A case study. *Amer. Statist.* **53** 89–93.

WILTFANG, J., ESSELMANN, H., BIBL, M., HÜLL, M., HAMPEL, H., KESSLER, H., FRÖLICH, L., SCHRÖDER, J., PETERS, O., JESSEN, F., LUCKHAUS, C., PERNECZKY, R., JAHN, H., FISZER, M., MALER, J. M., ZIMMERMANN, R., BRUCKMOSER, R., KORNHUBER, J. and LEWCZUK, P. (2007). Amyloid beta peptide ratio 42/40 but not A beta 42 correlates with phospho-Tau in patients with low- and high-CSF A beta 40 load. *J. Neurochem.* **101** 1053–1059.

YEE, T. W. (2015). *Vector Generalized Linear and Additive Models*: *With an Implementation in R*. Springer, New York. MR3408425

M. BERGER
DEPARTMENT OF MEDICAL BIOMETRY,
INFORMATICS AND EPIDEMIOLOGY
UNIVERSITY OF BONN/UNIVERSITY HOSPITAL BONN
SIGMUND-FREUD-STRASSE 25
D-53105 BONN
GERMANY
E-MAIL: Moritz.Berger@imbie.uni-bonn.de

M. WAGNER
DEPARTMENT OF PSYCHIATRY
AND PSYCHOTHERAPY
UNIVERSITY OF BONN
SIGMUND-FREUD-STRASSE 25
D-53105 BONN
GERMANY
AND
GERMAN CENTER FOR
NEURODEGENERATIVE DISEASES
SIGMUND-FREUD-STRASSE 25
D-53105 BONN
GERMANY
E-MAIL: Michael.Wagner@dzne.de

M. SCHMID
DEPARTMENT OF MEDICAL BIOMETRY,
INFORMATICS AND EPIDEMIOLOGY
UNIVERSITY OF BONN/UNIVERSITY HOSPITAL BONN
SIGMUND-FREUD-STRASSE 25
D-53105 BONN
GERMANY
AND
GERMAN CENTER FOR
NEURODEGENERATIVE DISEASES
SIGMUND-FREUD-STRASSE 25
D-53105 BONN
GERMANY
E-MAIL: Matthias.Schmid@imbie.uni-bonn.de