Published online 28 June 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 2 **1**
https://doi.org/10.1093/nargab/lqad062

Genome-wide association study of a semicontinuous trait: illustration of the impact of the modeling strategy through the study of Neutrophil Extracellular Traps levels

Gaëlle Munsch[1,*], Carole Proust[1], Sylvie Labrouche-Colomer[2,3], Dylan Aïssi[1], Anne Boland[4], Pierre-Emmanuel Morange[5], Anne Roche[6], Luc de Chaisemartin[7,8], Annie Harroche[9], Robert Olaso[4,10], Jean-François Deleuze[4,10], Chloé James[2,3], Joseph Emmerich[11], David M. Smadja[12,13], Hélène Jacqmin-Gadda[1,†] and David-Alexandre Trégouët[1,†]

[1]Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, F-33000 Bordeaux, France, [2]UMR1034, Inserm, Biology of Cardiovascular Diseases, University of Bordeaux, Pessac, France, [3]Laboratoire d'Hématologie, CHU de Bordeaux, Pessac, France, [4]Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), 91057 Evry, France, [5]Cardiovascular and Nutrition Research Center (C2VN), INSERM, INRAE, Aix-Marseille University, Marseille, France, [6]Service pneumologie hôpital Bicêtre, France, [7]Service Auto-immunité, Hypersensibilité et Biothérapies, Hôpital Bichat, Assistance Publique-Hôpitaux de Paris, Paris, France, [8]Université Paris-Saclay, INSERM, Inflammation, Microbiome, Immunosurveillance, Orsay, France, [9]Service d'Hématologie Clinique Centre de Traitement de l'Hémophilie Hôpital Necker Enfants Malades, France, [10]Centre d'Etude du Polymorphisme Humain, Fondation Jean Dausset, Paris, France, [11]Department of vascular medicine, Paris Saint-Joseph Hospital Group, University of Paris, UMR1153, INSERM, CRESS, 185 rue Raymond Losserand, Cité, 75674, France, [12]Innovative Therapies in Hemostasis, Université de Paris, INSERM, F-75006 Paris, France and [13]Hematology Department and Biosurgical Research Lab (Carpentier Foundation), Assistance Publique Hôpitaux de Paris, Centre-Université de Paris (APHP-CUP), F-75015 Paris, France

Received January 25, 2023; Revised May 10, 2023; Editorial Decision June 01, 2023; Accepted June 07, 2023

ABSTRACT

Over the last years, there has been a considerable expansion of genome-wide association studies (GWAS) for discovering biological pathways underlying pathological conditions or disease biomarkers. These GWAS are often limited to binary or quantitative traits analyzed through linear or logistic models, respectively. In some situations, the distribution of the outcome may require more complex modeling, such as when the outcome exhibits a semicontinuous distribution characterized by an excess of zero values followed by a non-negative and right-skewed distribution. We here investigate three different modeling for semicontinuous data: Tobit, Negative Binomial and Compound Poisson-Gamma. Using both simulated data and a real GWAS on Neutrophil Extracellular Traps (NETs), an emerging biomarker in immuno-thrombosis, we demonstrate that Compound Poisson-Gamma was the most robust model with respect to low allele frequencies and outliers. This model further identified the MIR155HG locus as significantly ($P = 1.4 \times 10^{-8}$) associated with NETs plasma levels in a sample of 657 participants, a locus recently highlighted to be involved in NETs formation in mice. This work highlights the importance of the modeling strategy for GWAS of a semicontinuous outcome and suggests Compound Poisson-Gamma as an elegant but neglected alternative to Negative Binomial for modeling semicontinuous outcome in the context of genomic investigations.

INTRODUCTION

Semicontinuous data, characterized by an excess of zeros followed by a non-negative and right-skewed distribution,

*To whom correspondence should be addressed. gaelle.munsch@u-bordeaux.fr
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

are frequently observed in biomedical research (1). When the study aims at identifying determinants of such a semicontinuous biomarker, it must be handled as the outcome variable and due to the inflation of zeros, classical models such as linear regression cannot be applied without violating the Gaussian assumptions, even with a logarithmic or rank-based inverse-normal transformation. For instance when the interest specifically lies in the identification of molecular determinants associated with a disease semicontinuous biomarker, as it is encountered in the omics era in order to identify/characterize new biological pathways, inform about drug discovery and help in individual risk prediction (2), the problem of how to model its distribution arises. Besides, in the context of Genome Wide Association Studies (GWAS), linear and logistic regression are often the only statistical models implemented in popular software. Users are then encouraged to transform their outcome of interest into a binary or a Gaussian variable at the cost of a loss of information and/or of complex interpretation of genetic association parameters.

Over the past decades several statistical models have been developed to model semicontinuous data by taking into account the mass of zeros. Among the most commonly used models are the Tobit and the two-part models (3,4).

The two-part model and its extensions (5,6) rely on the use of a logistic regression model to predict the probabilities of occurrence of zero values and of a linear regression model for the analysis of the strictly continuous outcome. The main assumption of this model is that the values of the outcome are derived from two different generating processes. This model has been used in various applications including the modeling of tumor size in cancer, food intake, microbiome abundance or individual costs of chronic kidney disease (7–11). However, the two-part model does not make possible the estimation of a single parameter that represents the association of an explanatory variable on the outcome. In contrast to the two-part model, Tobit models consider a single distribution of the outcome. In the case of zero-inflated data, the Tobit model assumes that the semicontinuous variable is a truncated observation of a Gaussian variable. This modeling is mainly used to account for floor or ceiling effect of the outcome variable that could be due to technical measurement limits (12–15).

Another possibility is to consider the outcome variable as quantitative discrete, which can be done in some cases by changing the unit of measurement through the use of a multiplicative factor, without losing precision. In this case, models for count data such as the Poisson model or the Negative Binomial model in presence of overdispersion can be used. These models are relevant as long as the proportion of zeros is not too high (16,17). As the Tobit model, these models allow for a simple interpretation of the results since only one coefficient is estimated per explanatory variable. Extensions of these models have been developed to account for the zero mass (also known as ZIP for Zero-Inflated Poisson and ZINB for Zero-Inflated Negative Binomial) but they make the assumption that the distribution of the outcome is composed of two generating processes, like the two-part models.

New models based on so-called Tweedie distributions (18–20) have recently emerged for the analysis of semicontinuous data but their use remains marginal (21). The Compound Poisson-Gamma model belongs to this Tweedie family. It assumes that the semicontinuous outcome is defined as a Poisson sum of gamma random variables. Semicontinuous data are then modeled through the use of a single distribution.

The choice between these different models is not obvious as each semicontinuous outcome has its own properties. As there is no established decision tool, the model to be applied should be chosen according to the distribution of the outcome and the clinical context (22).

In this work, we show the impact of the adopted model on the results of a GWAS that aimed at identifying genetic factors associated with plasma levels of Neutrophil Extracellular Traps (NETs), a semicontinuous biomarker involved in thrombosis. We highlight the differences between the models with respect to the flexibility of the underlying assumptions, the robustness to outliers and low allele frequency that can help to select the most appropriate model for future studies.

NETs are one of the emerging biomarkers with a key role in thrombosis that often present an excess of zero values (23–25). In the event of a vascular breach, neutrophils and platelets are the first cells to be recruited and activated (26). When neutrophils are activated by platelets, they have pro-inflammatory properties that can enhance tissue damage and induce thrombus formation in particular when they evolve towards a certain form of cell death leading to the release of their decondensed chromatin as a network of fibres also called NETs. NETs are composed of DNA fibres comprised of antimicrobial proteins and histones which promote coagulation, platelets activation and thus thrombus formation (27,28). NETs are involved in many other biological mechanisms such as immune response to viruses, diabetes, cystic fibrosis, cancer tumor growth, progression and metastasis (29–33).

NETs plasma levels were here measured in 657 participants of the ≪ FActeurs de RIsque et de récidives de la maladie thromboembolique Veineuse ≫ (FARIVE) study (34). Genome wide genotype data were also available for these participants and then used to conduct a GWAS on NETs levels. We illustrate how the results of this GWAS are impacted by the statistical approach adopted to model NETs plasma levels.

MATERIALS AND METHODS

The FARIVE study

The FARIVE study is a multicenter case-control study conducted between 2003 and 2007. The sample includes 607 patients with a documented episode of deep vein thrombosis and/or pulmonary embolism and 607 healthy individuals. A detailed description of the study can be found elsewhere (34). Briefly, patients were not eligible if they were younger than 18 years, had previous venous thrombosis (VT) event, active cancer or recent history of malignancy (within 5 years). Controls were recruited over the same period and matched to cases according to age and sex. They did not have any history of venous and arterial thrombotic disease as well as cancer, liver or kidney failure.

NETs measurements. NETs were quantified by measuring myeloperoxidase (MPO)-DNA complexes using an in-house capture ELISA already described (35) in a subsample of 410 VT patients (7 months after their inclusion in the study once the anticoagulant treatment has stopped) and 327 controls (at their time of inclusion in the study). Briefly, microtiter plates were coated with anti-human MPO antibody. After blocking, serum samples were added together with a peroxidase-labeled anti-DNA antibody. After incubation, the peroxidase substrate was added and absorbance measured at 405 nm in a spectrophotometer.

Genotyping and imputation. FARIVE participants were genotyped using the Illumina Infinium Global Screening Array v3.0 (GSAv3.0) microarray at the Centre National de Recherche en Génomique Humaine (CNRGH). Individuals with at least one of the following criteria were excluded: discordant sex information, relatedness individuals identified by pairwise clustering of identity by state distances (IBS), genotyping call rate lower than 99%, heterozygosity rate higher/lower than the average rate ±3 standard deviation or of non-European ancestry. After applying these criteria, the final sample was composed of 1077 individuals. Among the 730059 variants genotyped, 145238 variants without a valid annotation were excluded as well as 656 variants deviating from Hardy-Weinberg equilibrium in controls at $P < 10^{-6}$, 47 286 variants with a Minor Allele Count (MAC) lower than 20 and 1774 variants with a call rate lower than 95%. This quality controls procedure was conducted using Plink v1.9 (36) and the R software v3.6.2. Finally, there were 535105 markers left for imputation which was then performed with Minimac4 using the 1000 Genomes phase 3 version 5 reference panel (37).

Genome wide association study of NETs plasma levels. The present study relies on a subsample of 657 individuals (372 VT cases and 285 controls) with both NETs measurements and imputed genetic data. All genetic variants with minor allele frequency (MAF) greater than 0.01 and imputation quality score $r^2 > 0.3$ were tested for association with NETs plasma levels. As shown in the next section, three different statistical models were deployed. In all, associations were tested on imputed allelic doses and adjusted for potential confounders that is age, sex, smoking, case-control status and the four first principal components derived from genome wide genotype data (38–40). The standard genome-wide statistical threshold of 5×10^{-8} was used to consider genetic polymorphisms as significantly associated with NETs plasma levels.

Statistical modeling for GWAS analysis of NETs plasma levels

Since we were interested in identifying genetic factors that influence mean NETs plasma levels, any statistical approach that treats independently the zeros mass and the distribution of positive values, such as the two-part, ZIP and ZINB models, was deemed not adapted to our application. As a consequence, only three models were compared in this study: Tobit, Compound Poisson-Gamma and Negative Binomial models. The Poisson model was not investigated because NETs plasma data presented a large overdispersion (see Results section), a situation where Negative Binomial model is preferable. In this study, we aimed to identify the most suitable model for semicontinuous data in order to conduct a GWAS on NETs and highlight the difference between the three models.

Tobit model. In the Tobit model, the observed variable Y is assumed to be a right or left truncated observation of an underlying Gaussian latent variable (Y^*). Let c the constant threshold for truncation which needs to be known and is equal to zero in the context of zero-inflated data. Therefore, the Tobit model assumes that zero values are due to censoring or measurement limits and so they do not represent the true absence of the variable. In the case of a left truncation at 0 the values of the observed variable are:

$$Y = \begin{cases} Y^* \; if \; Y > 0 \\ 0 \; otherwise \end{cases}$$

The subsequent regression model is:

$$\mathbb{E}(Y^* | \; X) = \beta X$$

where $\mathbb{E}(Y^* | \; X)$ is the expected value of the underlying Gaussian variable Y^* conditioned on the explanatory variables X, and where β represents the regression parameters associated to X. The Tobit model is available in the *VGAM* R package (41).

Compound Poisson-Gamma model. An Exponential Dispersion Model (EDM) is a two-parameter family of distributions composed of a linear exponential family with an additional dispersion parameter (42). EDMs are characterized by their variance function $\mathbb{V}(.)$ that is an exponential function used to describe the relationship between the mean and the variance. If Y follows an EDM, then $\mathbb{E}(Y) = \mu$ and $Var(Y) = \Phi\mathbb{V}(\mu)$ with Φ a dispersion parameter. Tweedie models are a class of EDMs characterized by a power variance function: $\mathbb{V}(\mu) = \mu^p$ with p the index parameter (43,44). Most of the usual distributions are included in the class of Tweedie models such as the normal ($p = 0$), Poisson ($p = 1$), gamma ($p = 2$) and the inverse Gaussian ($p = 3$) (45).

The probability density function of a Tweedie model is defined as (42):

$$f(y|\mu, \; \Phi, \; p) = a(y, \; \Phi, \; p) \exp\left(\frac{1}{\phi}\left(y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right)\right)$$

where $a(.)$ is a given function.

The Compound Poisson-Gamma model belongs to the family of Tweedie models with $p \in]1; 2[$. It simultaneously models the occurrence and the intensity of the semicontinuous outcome (46). The distribution of a variable Y following a Compound Poisson-Gamma model may be defined as a Poisson sum of M Gamma distributions:

$$Y = \begin{cases} 0, \text{ if } M = 0 \\ K_1 + K_2 + \ldots + K_M, \text{ if } M > 0 \end{cases}$$

where $M \sim Pois(\lambda)$, $K_i \sim Gamma(\alpha, \gamma)$with α the shape parameter and γ the scale parameter, and where the values of K_i are iid and independent on M.

The Compound Poisson-Gamma model is a Tweedie model with the following parametrisation:

$$\mu = \lambda\alpha\gamma \; ; \Phi = \frac{\lambda^{1-p} * (\alpha\gamma)^{2-p}}{2-p} \; ; \; p = \frac{\alpha+2}{\alpha+1} \in]1;2[$$

$$\mathbb{E}(Y) = \mu = \lambda\alpha\gamma$$

$$Var(Y) = \Phi\mu^p = \lambda\gamma^2\alpha * (1+\alpha)$$

Thus, direct modeling of the global expectation $\mathbb{E}(Y)$ is possible using a generalized linear model with a logarithmic link function to insure positivity of the means:

$$log(\mathbb{E}(Y \mid X)) = \beta X$$

We used the *cplm* R package to implement Compound Poisson-Gamma models (47).

Negative binomial models. We also attempted to use a model for count data by multiplying NETs' values by 1000 to ensure discreteness without creating new ex-aequos. Let Y be a random variable following a Poisson distribution which depends on a single parameter $\lambda > 0$:

$$\mathbb{E}(Y) = Var(Y) = \lambda$$

The Poisson model is adapted to model the expectation of a count variable using a generalized linear model with a logarithmic link function:

$$log(\mathbb{E}(Y \mid X)) = \beta X$$

The Negative Binomial model is an extension to the Poisson model in the presence of over-dispersion of the outcome: $Var(Y) > \mathbb{E}(Y)$ (48,49). In that case, the variance of Y is linked to its expectation through the following relationship: $Var(Y) = \mathbb{E}(Y) + k * \mathbb{E}(Y)^2$ where $k > 0$ is a dispersion parameter. This model is also part of generalized linear models and its link function is the logarithm. Negative Binomial model can also be represented as Poisson distributions with a Gamma distributed means where $Y \sim Pois(\lambda)$ and $\lambda \sim Gamma(\alpha, \gamma)$ (50). However, unlike the Compound Poisson-Gamma presented above, the two variables Y and λ are not independent from each other.

Models' comparison

The three aforementioned tested models were applied to NETs data while adjusting for age, sex, smoking and case-control status. The fit of these models to NETs data were assessed in two ways. First, we computed the Root Mean Square Error (RMSE) of each tested model defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

where N is the sample size, $\hat{y}_i$ the prediction of the ith individual according to its covariates provided by a given modeling approach and y_i the observed value. Instead of predicting $\hat{y}_i$ by $\mathbb{E}(Y \mid X, \hat{\beta})$ that cannot be equal to zero, we used simulated predictions. For each studied model, a random value was generated for each individual according to its explanatory variables and the estimated model parameters. This process was repeated 1000 times and the mean of RMSEs over the 1000 replicates was reported. As the Tobit model predicts negative values that are not observed in our semicontinuous outcome, these were imputed at zero to calculate the corresponding RMSE.

Second, we graphically assessed the fit of models predictions using Quantile-Quantile plot (QQplot) of the observations and predictions for each tested model.

Simulation study

A simulation study was conducted to evaluate the control of the type I error (α) of the Negative Binomial and Compound Poisson-Gamma models in the context of genetic association studies as well as their sensibility to outliers. From the observed NETs data distribution, we randomly generated $S = 1, \ldots, 10000$ bootstrapped samples of size $N = 657$. For each bootstrapped sample, all individuals were randomly assigned four independent genotypes under the assumption of Hardy-Weinberg and corresponding to 4 genetic variants with allele frequencies 0.01, 0.05, 0.10 and 0.20. The association of genetic variants with the outcome was tested under the assumption of additive allele effects. This procedure was used to simulate semicontinuous data that mimic the NETs data observed in FARIVE and to allow the evaluation of the robustness of the two studied models (Negative Binomial and Compound Poisson-Gamma models) to a deviation from their underlying distribution. To assess the robustness to outliers, each simulated dataset was also analyzed after the exclusion of individuals with NETs level higher than 0.5, a threshold corresponding on average to the exclusion of 3% of individuals. In the FARIVE data, the proportion of zero was 15.8%. In a complementary simulation study, we assessed the robustness of the models by setting the proportion of zero values at 5%, 30%, 50% and 80%. To do this, we first fixed the number of zeros that the sample should contain and then we bootstrapped the strictly positive values of NETs from the FARIVE study for the remaining samples, so that the total sample size was $N = 657$ in each of the scenarios studied.

For each tested model, the number of times a genetic polymorphism was found statistically significant at $\alpha = 0.05$, $\alpha = 0.01$, and $\alpha = 0.001$ was used to compute its empirical type I error.

RESULTS

Population characteristics

The main characteristics of the FARIVE participants used in this work are presented in Table 1. There is approximately 40% of men, 20% of current smokers and individuals are on average 53 years old. The distribution of NETs plasma levels observed in FARIVE is shown in Figure 1A. Approximately 16% of exact zeros were observed with a higher proportion among VT cases compared to controls (20% and 10% respectively). To analyze NETs as count data, observed values were multiplied by 1000. This induced a large overdispersion (mean = 78; variance = 26 064) leading to the adoption of a Negative Binomial model for analysing such data.

Table 1. Main characteristics of the FARIVE study

	Total $N = 657$ N (%)	VT[b] cases $N = 372$ N (%)	Controls $N = 285$ N (%)
Sex—men	256 (39.0%)	141 (37.9%)	115 (40.4%)
Age at sampling (mean ± SD[a])	53.0 ± 18.8	53.3 ± 19.3	52.8 ± 18.2
Smoking status			
Current smoker	128 (19.5%)	62 (16.7%)	66 (23.2%)
Former smoker/never	529 (80.5%)	310 (83.3%)	219 (76.8%)
Neutrophil Extracellular Traps levels			
All values			
Mean ± SD	0.08 ± 0.16	0.05 ± 0.11	0.12 ± 0.20
Median [Q1;Q3]	0.03 [4×10^{-3};0.07]	0.03 [2×10^{-3};0.05]	0.04 [0.01;0.12]
Exact zero	104 (15.8%)	75 (20.2%)	29 (10.2%)

[a]Standard deviation.
[b]Venous thrombosis.

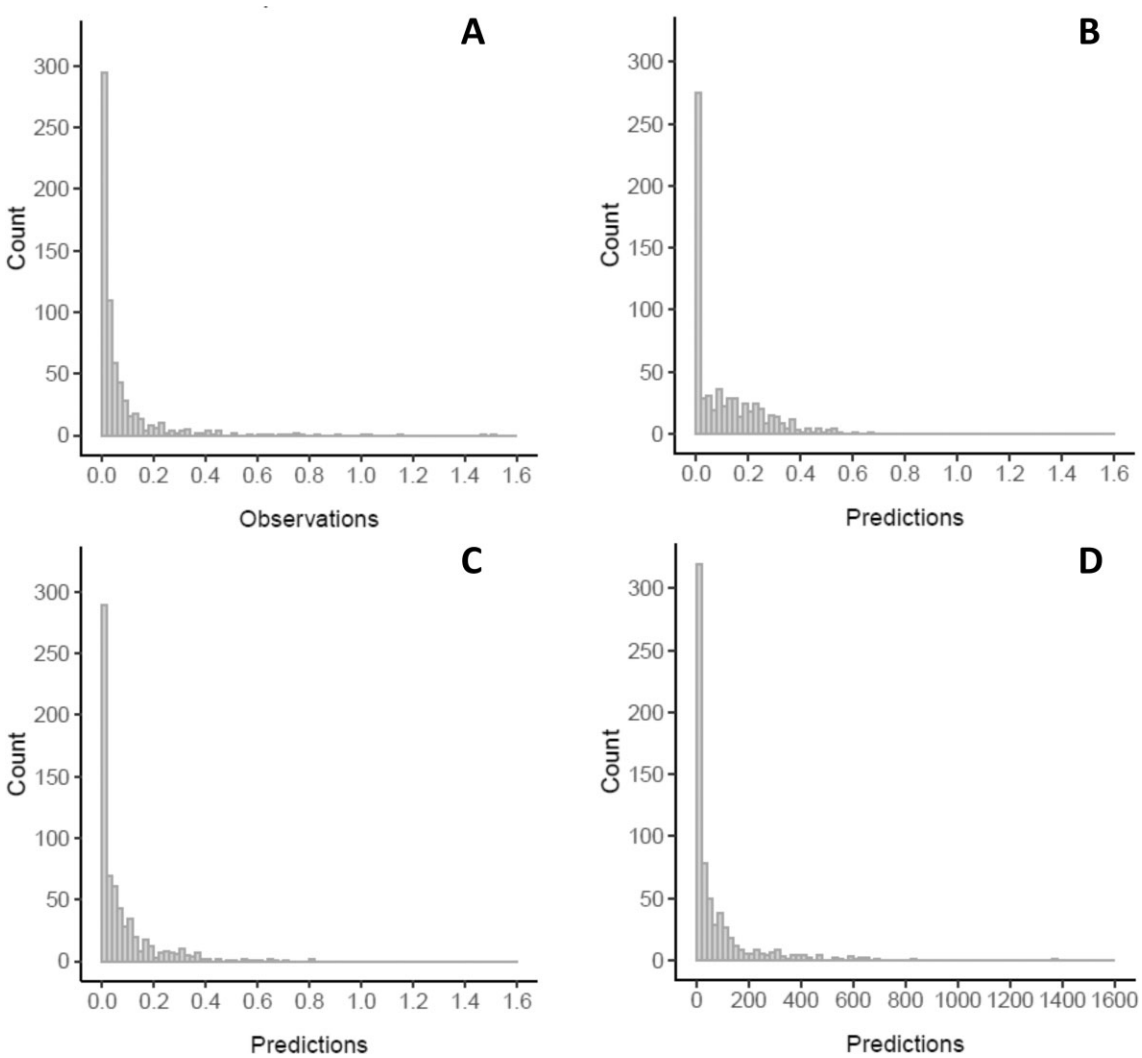

Figure 1. This figure presents the distribution of the observed NETs plasma levels (**A**), predictions from the Tobit model (**B**), the Compound Poisson-Gamma model (**C**) and the Negative Binomial model (**D**).

Table 2. Comparison of regression parameter estimates on the FARIVE data according to the three models

	Tobit Beta (SD)	Compound Poisson-Gamma Beta (SD)	Negative Binomial[a] Beta (SD)
Covariates			
Age (10 years)	0.005 (0.004)	0.05 (0.04)	0.05 (0.04)
Sex (males)	-0.01 (0.02)	-0.08 (0.16)	-0.04 (0.13)
Smoking (non-smokers)	0.05 (0.02)	0.50 (0.19)	0.52 (0.17)
Status (controls)	-0.08 (0.01)	-0.87 (0.15)	-0.87 (0.13)
RMSE[b]	198.7[c][145.5; 251.9]	193.9 [180.8; 207.1]	211.5 [187.0; 236.0]

[a]For the distribution of NETs multiplied by 1000.
[b]Mean [min–max] over 1000 bootstrapped samples.
[c]Negative predictions were censored at zero.

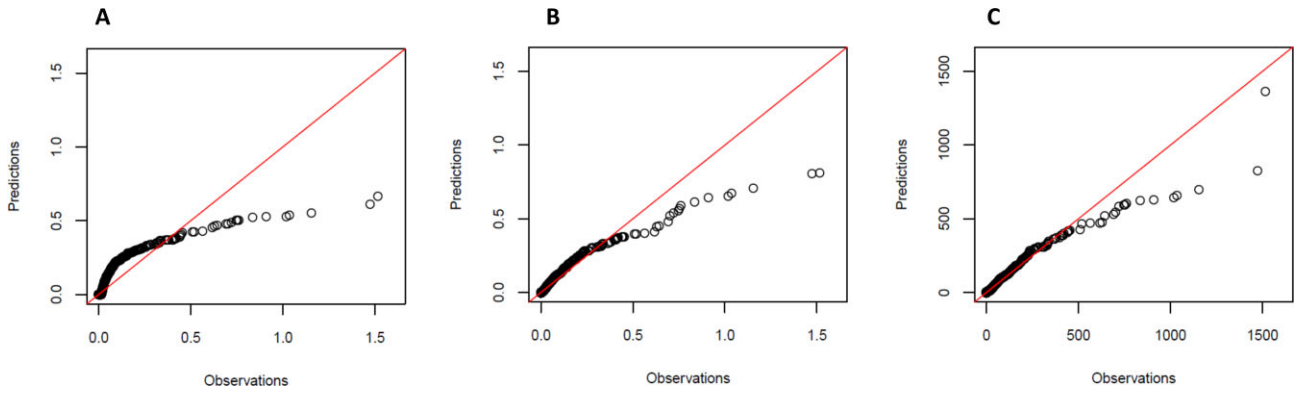

Figure 2. This figure presents the Quantile-Quantile plots of observations and predictions from Tobit (**A**), Compound Poisson-Gamma (**B**) and Negative Binomial (**C**) models. The red line represents the perfect match between observations and predictions.

Clinical variables & goodness of fit

Table 2 reports the association of clinical covariates with NETs plasma levels in each of the three studied models, Tobit Compound Poisson-Gamma and Negative Binomial. The Tobit model assumes a linear association of the covariates on the expected mean of the latent variable, i.e. the true value of NETs. For example, each 10-year increase in age is associated with an increase of 0.005 on the expected mean of the latent variable of NETs plasma levels, given the other covariates are held constant. Regarding the two other models, as a logarithmic link function is used, the association of covariates on the expected mean of NETs is multiplicative. As a consequence, for the Compound Poisson-Gamma model, each 10-year increase of age is associated with an expected mean of NETs plasma levels multiplied by 1.05 ($= e^{0.05}$). Similar interpretation holds for the Negative Binomial model that yielded regression parameters very close to those obtained via the Compound Poisson-Gamma models.

RMSEs provided by the three models are shown in Table 2. The lowest RMSE was observed for the Compound Poisson-Gamma model while the Negative Binomial model exhibited the highest one. Graphically, the Compound Poisson-Gamma (Figure 1C) and the Negative Binomial (Figure 1D) showed similar distributions of their predicted values even if the right skewedness was slightly less pronounced for the Compound Poisson-Gamma distribution. These distributions were rather close to that observed for the original NETs data (Figure 1A). By contrast, the Tobit distribution (Figure 1B) substantially deviated from the original data and looked like a left-truncated Gaussian distribution.

Quantile-Quantile plots of the observed versus predicted values did not visually show obvious deviation from the bisection line, except for high values (>0.5 in the original NETs scale), for the Compound Poisson-Gamma (Figure 2B) and the Negative Binomial (Figure 2C) models. Conversely, for the Tobit model (Figure 2A), the QQplot line deviated from the bisection line from the lower values.

Altogether, these observations suggest that the Compound Poisson-Gamma model seems the most adequate to analyze FARIVE NETs data. Nevertheless, we conducted a GWAS on NETs plasma levels using each of the three models discussed above in order to get additional elements of comparison between these models.

GWAS analysis on NETs plasma levels

A total of 9 670 724 autosomal genetic variants with imputation criterion $r^2 > 0.3$ and minor allele frequencies (MAF) > 0.01 were tested for association with NETs plasma levels using the Tobit, Negative Binomial and Compound Poisson-Gamma models. Quantile-Quantile plots for the observed and expected p-values summarizing the GWAS results for each model are shown in Figure 3. While the whole set of association results was compatible with what was expected under the null hypothesis of no genetic association for the Compound Poisson-Gamma model (Figure 3B), strong deviations were observed for the Tobit (Figure 3A) and Negative Binomial (Figure 3C) models. By restrict-

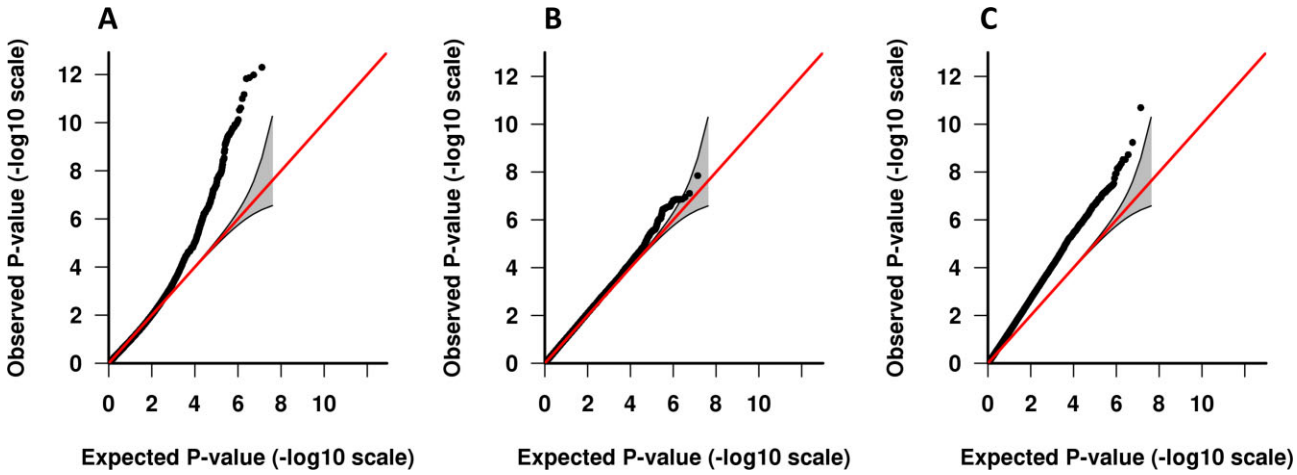

Figure 3. This figure presents the Quantile–Quantile plots of the *P*-values distributions from the GWAS with Tobit (**A**), Compound Poisson-Gamma (**B**) and Negative Binomial (**C**) models.

ing the GWAS results to genetic polymorphisms with MAF >5%, inflation was no longer observed for the Tobit model (Supplementary Figure S1A, genomic inflation factor $\lambda = 0.96$) while the Negative Binomial model remained strongly inflated (Supplementary Figure S1C, $\lambda = 1.46$).

To further explore the remaining inflation, we re-ran the GWAS under the Compound Poisson-Gamma and Negative Binomial models after excluding 19 FARIVE participants (~3%) with NETs plasma levels higher than 0.5. Inflation in the Negative Binomial model was considerably decreased (Supplementary Figure S2B) and completely vanished when we additionally restricted the GWAS analysis on genetic variants with MAF >5% (Supplementary Figure S3B, $\lambda = 1.03$).

Finally, we conducted simulation studies (see Methods) that confirmed that the type I error of the association test in the Compound Poisson-Gamma model was controlled whereas the association test in the Negative Binomial model exhibited inflated type I error (α) for the three nominal values of α considered (Supplementary Table S1) when data distribution fit the one observed for NETs plasma levels in the FARIVE study. Type I error was rather well controlled in absence of extreme values (Supplementary Table S2). These conclusions also hold in presence of strong linkage disequilibrium between tested variants (Supplementary Text, Supplement Table S3, Supplementary Figure S4). Compound Poisson-Gamma model was also robust when the proportion of zero values varied (5%, 30%, 50% and 80% of zero) (Supplementary Tables S4–S7). Simulation results showed that the type I error was controlled in all settings for the Compound Poisson-Gamma model, whereas the type I error was inflated for the Negative Binomial model when the zeros represent less than 50% of the distribution.

All these observations add support for the use of the Compound Poisson-Gamma model for the GWAS analysis of NETs plasma levels.

The corresponding Manhattan plot shown in Figure 4 revealed one genome-wide significant locus. The lead polymorphism rs57502213 is a deletion of two nucleotides (TC), mapping to the miR-155 hosting gene (*MIR155HG*). This variant had a MAF of ~7%, exhibited a good imputation quality ($r^2 = 0.92$) and its minor allele was associated with a 2.53-fold increase (95% confidence interval [1.85–3.47], $P = 1.42 \times 10^{-8}$) in NETs plasma levels. The average NETs plasma levels were higher in carriers of the deletion of the TC allele than in non-carriers (0.15 versus 0.07). This pattern of association was consistent in VT cases and in controls (Table 3). Furthermore, this association was robust to the exclusion of the single homozygous individual for the two base pair deletion (2.58 [1.87–3.56] $P = 1.4 \times 10^{-8}$).

Table 3. Association of rs57502213 with NETs plasma levels in the FARIVE study

	Genotype for rs57502213		
	TC/TC	TC/–	–/–
All individuals (*N* = 657)			
N	568	88	1
Mean ± SD	0.07 ± 0.13	0.15 ± 0.28	0.07
Exact zero	91 (16.0%)	13 (14.8%)	-
Cases (*N* = 372)			
N	323	48	1
Mean ± SD	0.04 ± 0.06	0.12 ± 0.26	0.07
Exact zero	66 (20.4%)	9 (18.7%)	-
Controls (*N* = 285)			
N	245	40	-
Mean ± SD	0.10 ± 0.18	0.20 ± 0.30	-
Exact zero	25 (10.2%)	4 (10.0%)	-

As shown in the locus zoom of the region provided in Supplementary Figure S5, other variants with varying degrees of linkage disequilibrium with the top rs57502213 were also highly associated with NETs plasma levels in the FARIVE study. In order to explore the signal obtained in this region and especially variants close to the *GABPA* and *APP* genes, good candidates according to the literature (51–53) and otherwise implicated in the development of neurodegenerative diseases, we further conducted a complementary haplotype association using the *haplo.stats R package* (54). Based on the linkage disequi-

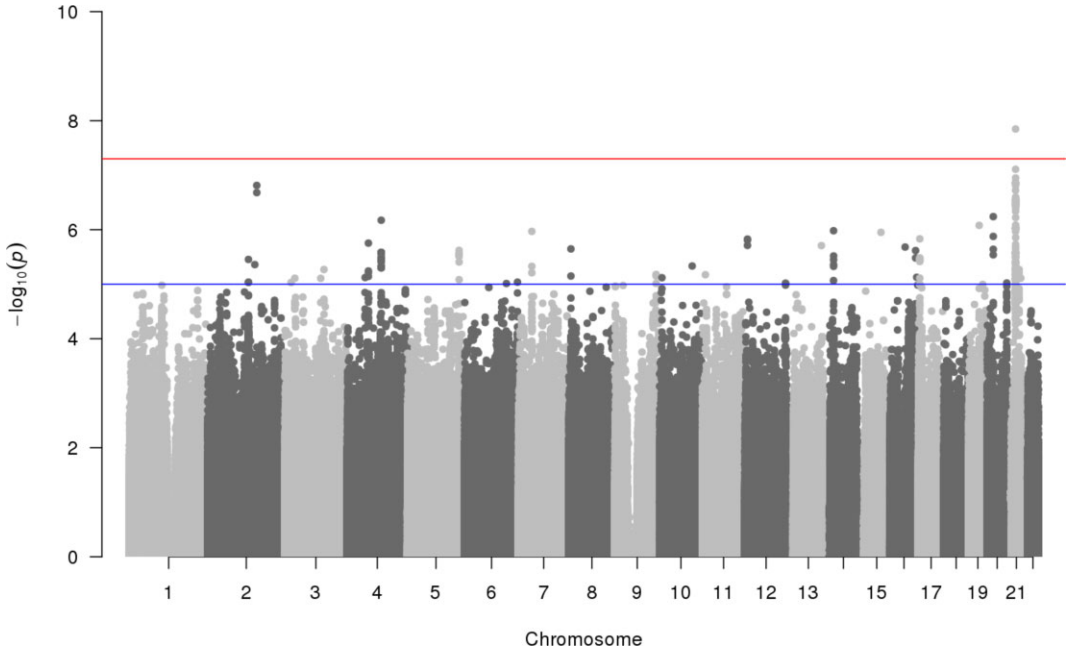

Figure 4. The $-log_{10}$ of the p-values are presented according to the position of the associated tested SNP across the genome. The genome wide significant threshold (5×10^{-8}) is represented with a red line.

librium and existing haplotypes between the most associated genetic variants ($P < 10^{-6}$) with NETs plasma levels in the GWAS, we identified five frequent haplotypes (H_1–H_5) which can be inferred with three polymorphisms as illustrated in Table 4. The association between these five haplotypes (under the assumption of additive effects) and NETs plasma levels was studied using the Compound Poisson-Gamma model and the same covariates as for GWAS. This analysis showed that all haplotypes carrying the rs57502213-A allele tended be associated with increased NETs plasma levels, but the main statistical signal was observed for the H_5 haplotype that was associated with a 3.91-fold increase (95% confidence interval [2.53–6.04], $P = 1.3 \times 10^{-9}$) in NETs plasma levels compared with the most common H_1 haplotype. This H_5 haplotype was the haplotype carrying the minor alleles of each of the three tested variants.

Full GWAS summary statistics are available on GWAS catalog under the accession number GCST90137414 (55).

DISCUSSION

This work was motivated by the search of genetic factors associated with NETs plasma levels exhibiting a semicontinuous distribution. In the literature, it is common to use transformation (i.e. rank-based inverse normal transformation) in order to normalize the phenotype of interest. Popular GWAS softwares such as REGENIE (56) have implemented this method and are often used, especially when GWAS on several phenotypes are performed. However, this transformation is not adapted to the presence of ex-aequos, and thus to semicontinuous distributions, and its utility has previously been discussed (57,58). A rank-based inverse normal transformation has been applied to NETs distribution but as shown in Supplementary Figure S6, it did not normalise NETs distribution because of the exact zeros and was therefore considered unsuitable for a GWAS with a linear model.

We here compared three different modeling strategies, Tobit, Negative Binomial and Compound Poisson-Gamma models, that handle the excess of zero and the asymmetric distribution while allowing the estimation of a single regression parameter for characterizing the association between an explanatory variable and the global mean of the semicontinuous outcome.

Visual inspection showed that both the Negative Binomial and Compound Poisson-Gamma models fit better the observed NETs distribution than the Tobit model. Indeed, the underlying hypothesis of a left-truncated Gaussian distribution with only two parameters makes the Tobit model less-flexible than Compound Poisson-Gamma and Negative Binomial models. RMSE analysis provided further support for the use of Compound Poisson-Gamma model. Of note, the definition of this model matches quite well the biological mechanisms underlying NETs production as it is intuitively reasonable to assume that the number of dead neutrophils follows a Poisson distribution, and that each of these rejects a certain quantity of NETs that would follow a Gamma distribution.

Our GWAS and simulation studies revealed that the Tobit and Negative Binomial models were prone to strong inflation of *P*-values. While this inflation could be attributable to genetic variants with low allele frequency (MAF < 5%) for the Tobit model, this inflation was due to both low allele frequency genetic polymorphisms and extreme positive values of the outcome for the Negative Binomial model. The poor control of type I error by the Negative Binomial model has already been highlighted in previous work

Table 4. Haplotype structure at the chr21q21.3 locus and its association with NETs plasma levels in the FARIVE study

				Haplotypes		
Haplotypes	rs73156700	rs57502213	rs35033826	Frequencies	RR [95%CI]	*P*
H_1	T	ATC	A	0.901	Reference	-
H_2	T	ATC	C	0.023	1.04 [0.59; 1.83]	0.89
H_3	T	A	A	0.018	1.37 [0.72; 2.59]	0.33
H_4	A	A	A	0.022	1.37 [0.77; 2.45]	0.28
H_5	A	A	C	0.026	3.91 [2.53; 6.04]	1.3×10^{-9}

Association was tested using a Compound Poisson Gamma model adjusted for age, sex, smoking, case–control status and genetically derived principal components.

in the context of RNA-Seq analysis (59). The Compound Poisson-Gamma model was much more robust to these two phenomena. Using Compound Poisson-Gamma model, we identified one significant locus on chr21q21.3 associated with NETs plasma levels. This locus maps to a long non coding RNA that hosts miR-155 (and as such is referred to as *miR155HG*, for Hosting Genes) and the lead polymorphism was rs57502213, an intronic deletion in *miR155HG*. While several recent publications have highlighted the role of miR-155 in the NETs formation (60,61), little information is available in public resources about the possible functional impact of rs57502213. This genetic variant is in moderate linkage disequilibrium (pairwise D' > 0.50) with several other nearby variants located in *MRPL39*, *GABPA* and *APP*, the latter having also been reported to be involved in NETs formation (51). Note that another GWAS on NETs plasma levels has recently been conducted in the Rotterdam study (62). Despite the large sample size of this study (∼5600 individuals), no significant genome-wide association was detected and the association of our lead polymorphism did not replicate there ($P = 0.14$). However, different kits were used to measure NETs levels in the two studies and recent works have emphasized the need for standardized methods for NETs measurements (63,64). Of note, in the Rotterdam study, NETs were analyzed using a log-transformed model suggesting that no zero values (or few) were observed (or were discarded). This contrasts with FARIVE data and could contribute to the heterogeneity of findings between studies. Nevertheless, because of the increasingly recognized role of the chr21q21.3 locus in NETs biology, further works deserve to be conducted to clarify the genetic signal observed in the present study.

To conclude, our work indicates that the modeling strategy for a semicontinuous outcome is crucial, but not straightforward. The choice of the model should take into account the nature of the (biological) process generating zero values, the distribution of the outcome and, especially, the presence of extreme values. The Tobit model with only two parameters is less flexible than Compound Poisson-Gamma and Negative Binomial models and our work shows that the Compound Poisson-Gamma model, while still marginally used, is more robust than the Negative Binomial model to outliers and low allele frequency making. This make it well suitable for GWAS analysis on semicontinuous trait, even in presence of related individuals where a Compound Poisson-Gamma mixed modeling is implemented in the *R package cplm.* The use of the Compound Poisson-Gamma model as an alternative to Negative Binomial model would also deserve to be explored in the context of RNA-Seq analysis.

DATA AVAILABILITY

Full GWAS summary statistics are available on GWAS catalog under the accession number GCP000431. The code supporting the current study is available on Zenodo 10.5281/zenodo.8006805.

ETHICS APPROVAL

Research have been performed in accordance with the Declaration of Helsinki. The FARIVE study was approved by the 'Comité consultatif de protection des personnes dans la recherche biomedicale' (Project no. 2002-034). Written informed consent to participate was obtained from all FARIVE participants.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

G.M. benefited from the EUR DPH, a PhD program supported within the framework of the PIA3 (Investment for the future). Project reference 17-EURE-0019.
Statistical analyses benefited from the CBiB computing centre of the University of Bordeaux.
This project was carried out in the framework of the INSERM GOLD Cross-Cutting program (P.-E.M., D.-A.T.) and of the French National Research Agency (ANR) ANR-18-RHUS-002 program as part of the French Investment for the Future project.

FUNDING

G.M. and D.-A.T. are supported by the EPIDEMIOM-VT Senior Chair from the University of Bordeaux initiative of excellence IdEX; Fondation pour la Recherche Médicale; Programme Hospitalier de recherche Clinique [PHRC 20 002, PHRC2009 RENOVA-TV]; Fondation de France; Leducq Foundation; FARIVE genetic data were funded by the GENMED Laboratory of Excellence on Medical Genomics [ANR-10-LABX-0013]; National Research Agency (ANR) as part of the French Investment for the Future.
Conflict of interest statement. None declared.

REFERENCES

1. Min,Y. and Agresti,A. (2002) Modeling Nonnegative Data with Clumping at Zero: a Survey. *JIRSS*, **1**, 7–33.
2. Folkersen,L., Gustafsson,S., Wang,Q., Hansen,D.H., Hedman,Å.K., Schork,A., Page,K., Zhernakova,D.V., Wu,Y., Peters,J. *et al.* (2020) Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.*, **2**, 1135–1148.
3. Cragg,J.G. (1971) Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, **39**, 829–844.
4. Tobin,J. (1958) Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24–36.
5. Farewell,V.T., Long,D.L., Tom,B.D.M., Yiu,S. and Su,L. (2017) Two-part and related regression models for longitudinal data. *Annu. Rev. Stat. Appl.*, **4**, 283–315.
6. Feng,X., Lu,B., Song,X. and Ma,S. (2019) Financial literacy and household finances: a Bayesian two-part latent variable modeling approach. *J. Empir. Finance*, **51**, 119–137.
7. Chen,E.Z. and Li,H. (2016) A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, **32**, 2611–2617.
8. Garbutt,D.J., Stern,R.D., Dennett,M.D. and Elston,J. (1981) A comparison of the rainfall climate of eleven places in West Africa using a two-part model for daily rainfall. *Arch. Meteorol. Geophys. Bioclimatol. Ser. B*, **29**, 137–155.
9. Hartman,B., Larson,C., Kunkel,C., Wight,C. and Warr,R.L. (2023) A two-part model of the individual costs of chronic kidney disease. *North Am. Actuar. J.*, https://doi.org/10.1080/10920277.2023.2177676.
10. Rustand,D., Briollais,L., Tournigand,C. and Rondeau,V. (2022) Two-part joint model for a longitudinal semicontinuous marker and a terminal event with application to metastatic colorectal cancer data. *Biostatistics*, **23**, 50–68.
11. Tooze,J.A., Midthune,D., Dodd,K.W., Freedman,L.S., Krebs-Smith,S.M., Subar,A.F., Guenther,P.M., Carroll,R.J. and Kipnis,V. (2006) A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *J. Am. Diet. Assoc.*, **106**, 1575–1587.
12. Amore,M.D. and Murtinu,S. (2021) Tobit models in strategy research: critical issues and applications. *Glob. Strategy J.*, **11**, 331–355.
13. Chen,T., Ma,S., Kobie,J., Rosenberg,A., Sanz,I. and Liang,H. (2016) Identification of significant B cell associations with undetected observations using a Tobit model. *Stat. Interface*, **9**, 79–91.
14. Debnath,A.K., Blackman,R. and Haworth,N. (2014) A Tobit model for analyzing speed limit compliance in work zones. *Saf. Sci.*, **70**, 367–377.
15. McBee,M. (2010) Modeling outcomes with floor or ceiling effects: an introduction to the Tobit model. *Gift. Child Q.*, **54**, 314–320.
16. van den Broek,J. (1995) A score test for zero inflation in a Poisson distribution. *Biometrics*, **51**, 738–743.
17. Allison,P.D. (2012) *Logistic Regression Using SAS: Theory and Application*. 2nd edn., SAS Institute.
18. Tweedie,M.C.K. (1984) An index which distinguishes between some important exponential families. Statistics: Applications and New Directions. In: Ghosh,J.K. and Roy,J. (eds.) *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*. Indian Statistical Institute, Calcutta, pp. 579–604.
19. Gilchrist,R. and Drinkwater,D. (2000) The use of the Tweedie distribution in statistical modelling. In: Bethlehem,J.G. and van der Heijden,P.G.M. (eds.) *COMPSTAT*. Physica-Verlag HD, Heidelberg, pp. 313–318.
20. Jørgensen,B., Martínez,J.R. and Vinogradov,V. (2009) Domains of attraction to Tweedie distributions. *Lith. Math. J.*, **49**, 399–425.
21. Kurz,C.F. (2017) Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Med. Res. Methodol.*, **17**, 171.
22. Brown,J.E. and Dunn,P.K. (2011) Comparisons of Tobit, linear, and Poisson-Gamma regression models: an application of time use data. *Sociol. Methods Res.*, **40**, 511–535.
23. Kimball,A.S., Obi,A.T., Diaz,J.A. and Henke,P.K. (2016) The emerging role of NETs in venous thrombosis and immunothrombosis. *Front. Immunol.*, **7**, 236.
24. de Boer,O., Li,X., Teeling,P., Mackaay,C., Ploegmakers,H., van der Loos,C., Daemen,M., de Winter,R. and van der Wal,A. (2013) Neutrophils, neutrophil extracellular traps and interleukin-17 associate with the organisation of thrombi in acute myocardial infarction. *Thromb. Haemost.*, **109**, 290–297.
25. Hisada,Y., Grover,S.P., Maqsood,A., Houston,R., Ay,C., Noubouossie,D.F., Cooley,B.C., Wallén,H., Key,N.S., Thålin,C. *et al.* (2020) Neutrophils and neutrophil extracellular traps enhance venous thrombosis in mice bearing human pancreatic tumors. *Haematologica*, **105**, 218–225.
26. Ruhnau,J., Schulze,J., Dressel,A. and Vogelgesang,A. (2017) Thrombosis, neuroinflammation, and poststroke infection: the multifaceted role of neutrophils in stroke. *J. Immunol. Res.*, **2017**, 5140679.
27. Laridan,E., Martinod,K. and De Meyer,S.F. (2019) Neutrophil extracellular traps in arterial and venous thrombosis. *Semin. Thromb. Hemost.*, **45**, 86–93.
28. Diamond,S.L. (2016) Systems analysis of thrombus formation. *Circ. Res.*, **118**, 1348–1362.
29. Ng,H., Havervall,S., Rosell,A., Aguilera,K., Parv,K., von Meijenfeldt,F.A., Lisman,T., Mackman,N., Thålin,C. and Phillipson,M. (2021) Circulating markers of neutrophil extracellular traps are of prognostic value in patients with COVID-19. *Arterioscler. Thromb. Vasc. Biol.*, **41**, 988–994.
30. Wang,L., Zhou,X., Yin,Y., Mai,Y., Wang,D. and Zhang,X. (2019) Hyperglycemia induces neutrophil extracellular traps formation through an NADPH oxidase-dependent pathway in diabetic retinopathy. *Front. Immunol.*, **9**, 3076.
31. Zhu,L., Liu,L., Zhang,Y., Pu,L., Liu,J., Li,X., Chen,Z., Hao,Y., Wang,B., Han,J. *et al.* (2018) High level of neutrophil extracellular traps correlates with poor prognosis of severe influenza A infection. *J. Infect. Dis.*, **217**, 428–437.
32. Martínez-Alemán,S.R., Campos-García,L., Palma-Nicolas,J.P., Hernández-Bello,R., González,G.M. and Sánchez-González,A. (2017) Understanding the entanglement: neutrophil extracellular traps (NETs) in cystic fibrosis. *Front. Cell. Infect. Microbiol.*, **7**, 104.
33. Masucci,M.T., Minopoli,M., Del Vecchio,S. and Carriero,M.V. (2020) The emerging role of neutrophil extracellular traps (NETs) in tumor progression and metastasis. *Front. Immunol.*, **11**, 1749.
34. Trégouët,D.-A., Heath,S., Saut,N., Biron-Andreani,C., Schved,J.-F., Pernod,G., Galan,P., Drouet,L., Zelenika,D., Juhan-Vague,I. *et al.* (2009) Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. *Blood*, **113**, 5298–5303.
35. Granger,V., Peyneau,M., Chollet-Martin,S. and de Chaisemartin,L. (2019) Neutrophil extracellular traps in autoimmunity and allergy: immune complexes at work. *Front. Immunol.*, **10**, 2824.
36. Chang,C.C., Chow,C.C., Tellier,L.C., Vattikuti,S., Purcell,S.M. and Lee,J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.
37. Das,S., Forer,L., Schönherr,S., Sidore,C., Locke,A.E., Kwong,A., Vrieze,S.I., Chew,E.Y., Levy,S., McGue,M. *et al.* (2016) Next-generation genotype imputation service and methods. *Nat. Genet.*, **48**, 1284–1287.
38. White,P.C., Hirschfeld,J., Milward,M.R., Cooper,P.R., Wright,H.J., Matthews,J.B. and Chapple,I.L.c. (2018) Cigarette smoke modifies neutrophil chemotaxis, neutrophil extracellular trap formation and inflammatory response-related gene expression. *J. Periodontal Res.*, **53**, 525–535.
39. Ortmann,W. and Kolaczkowska,E. (2018) Age is the work of art? Impact of neutrophil and organism age on neutrophil extracellular trap formation. *Cell Tissue Res.*, **371**, 473–488.
40. Gupta,S., Nakabo,S., Blanco,L.P., O'Neil,L.J., Wigerblad,G., Goel,R.R., Mistry,P., Jiang,K., Carmona-Rivera,C., Chan,D.W. *et al.* (2020) Sex differences in neutrophil biology modulate response to type I interferons and immunometabolism. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 16481–16491.
41. Yee,T.W. (2015) *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer, New York, NY.
42. Zhou,H., Qian,W. and Yang,Y. (2022) Tweedie gradient boosting for extremely unbalanced zero-inflated data. *Commun. Stat. - Simul. Comput.*, **51**, 5507–5529.
43. Jørgensen,B. (1987) Exponential dispersion models. *J. R. Stat. Soc. Ser. B Methodol.*, **49**, 127–145.

44. Fox,J. (2016) *Applied Regression Analysis and Generalized Linear Models*. T3rd edn., SAGE, Los Angeles.
45. Dunn,P.K. and Smyth,G.K. (2008) Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. *Stat. Comput.*, **18**, 73–86.
46. Dzupire,N.C., Ngare,P. and Odongo,L. (2018) A Poisson-Gamma Model for zero inflated rainfall data. *J. Probab. Stat.*, **2018**, 1–12.
47. Zhang,Y. (2013) Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models. *Stat. Comput.*, **23**, 743–757.
48. Hoef,J.M.V. and Boveng,P.L. (2007) Quasi-Poisson vs. Negative Binomial Regression: how should we model overdispersed count data?*Ecology*, **88**, 2766–2772.
49. Gardner,W., Mulvey,E.P. and Shaw,E.C. (1995) Regression analyses of counts and rates: poisson, overdispersed Poisson, and negative binomial models. *Psychol. Bull.*, **118**, 392–404.
50. Gorshenin,A.K. and Korolev,V.Y.. (2018) Scale mixtures of Frechet distributions as asymptotic approximations of extreme precipitation. *J. Math. Sci.*, **234**, 886–903.
51. Canobbio,I., Visconte,C., Momi,S., Guidetti,G.F., Zarà,M., Canino,J., Falcinelli,E., Gresele,P. and Torti,M. (2017) Platelet amyloid precursor protein is a modulator of venous thromboembolism in mice. *Blood*, **130**, 527–536.
52. Perdomo-Sabogal,A., Nowick,K., Piccini,I., Sudbrak,R., Lehrach,H., Yaspo,M.-L., Warnatz,H.-J. and Querfurth,R. (2016) Human lineage-specific transcriptional regulation through GA-binding protein transcription factor Alpha (GABPa). *Mol. Biol. Evol.*, **33**, 1231–1244.
53. Schmechel,D.E., Goldgaber,D., Burkhart,D.S., Gilbert,J.R., Gajdusek,D.C. and Roses,A.D. (1988) Cellular localization of messenger RNA encoding amyloid-beta-protein in normal tissue and in Alzheimer disease. *Alzheimer Dis. Assoc. Disord.*, **2**, 96–111.
54. Schaid,D.J., Rowland,C.M., Tines,D.E., Jacobson,R.M. and Poland,G.A. (2002) Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous. *Am. J. Hum. Genet.*, **70**, 425–434.
55. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
56. Mbatchou,J., Barnard,L., Backman,J., Marcketta,A., Kosmicki,J.A., Ziyatdinov,A., Benner,C., O'Dushlaine,C., Barber,M., Boutkov,B. *et al.* (2021) Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.*, **53**, 1097–1103.
57. McCaw,Z.R., Lane,J.M., Saxena,R., Redline,S. and Lin,X. (2020) Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, **76**, 1262–1272.
58. Beasley,T.M., Erickson,S. and Allison,D.B. (2009) Rank-based inverse normal transformations are increasingly used, but are they merited?*Behav. Genet.*, **39**, 580–595.
59. Gauthier,M., Agniel,D., Thiébaut,R. and Hejblum,B.P. (2020) dearseq: a variance component score test for RNA-seq differential analysis that effectively controls the false discovery rate. *NAR Genomics Bioinforma*, **2**, lqaa093.
60. Hawez,A., Taha,D., Algaber,A., Madhi,R., Rahman,M. and Thorlacius,H. (2022) MiR-155 regulates neutrophil extracellular trap formation and lung injury in abdominal sepsis. *J. Leukoc. Biol.*, **111**, 391–400.
61. Hawez,A., Al-Haidari,A., Madhi,R., Rahman,M. and Thorlacius,H. (2019) MiR-155 regulates PAD4-dependent formation of neutrophil extracellular traps. *Front. Immunol.*, **10**, 2462.
62. Donkel,S.J., Portilla Fernández,E., Ahmad,S., Rivadeneira,F., van Rooij,F.J.A., Ikram,M.A., Leebeek,F.W.G., de Maat,M.P.M. and Ghanbari,M. (2021) Common and rare variants genetic association analysis of circulating neutrophil extracellular traps. *Front. Immunol.*, **12**, 615527.
63. Prével,R., Dupont,A., Labrouche-Colomer,S., Garcia,G., Dewitte,A., Rauch,A., Goutay,J., Caplan,M., Jozefowicz,E., Lanoix,J.-P. *et al.* (2022) Plasma markers of neutrophil extracellular trap are linked to survival but not to pulmonary embolism in COVID-19-related ARDS patients. *Front. Immunol.*, **13**, 851497.
64. Rada,B. (2019) Neutrophil extracellular traps. *Methods Mol. Biol.*, **1982**, 517–528.