Received: 14 May 2024 | Revised: 18 June 2024 | Accepted: 26 June 2024

DOI: 10.1002/psp4.13207

TUTORIAL

Automated Poisson regression exposure–response analysis for binary outcomes with *PoissonERM*

Yuchen Wang[1] | **Luke Fostvedt**[2] | **Jessica Wojciechowski**[3] | **Donald Irby**[4] | **Timothy Nicholas**[3]

[1] Pfizer Inc, South San Francisco, California, USA

[2] Pfizer Inc, Cambridge, Massachusetts, USA

[3] Pfizer Inc, Groton, Connecticut, USA

[4] Pfizer Inc, San Diego, California, USA

Correspondence
Yuchen Wang, Pfizer Inc. 181 Oyster Point Blvd, South San Francisco, CA 94080, USA.
Email: yuchen.wang2@pfizer.com

Abstract

PoissonERM is an R package used to conduct exposure–response (ER) analysis on binary outcomes for establishing the relationship between exposure and the occurrence of adverse events (AE). While Poisson regression could be implemented with *glm(), PoissonERM* provides a simple way to semi-automate the entire analysis and generate an abbreviated report as an R markdown (Rmd) file that includes the essential analysis details with brief conclusions. *PoissonERM* processes the provided data set using the information from the user's control script and generates summary tables/figures for the exposure metrics, covariates, and event counts of each endpoint (each type of AE). After checking the incidence rate of each AE, the correlation, and missing values in each covariate, an exposure–response model is developed for each endpoint based on the provided specifications. *PoissonERM* has the flexibility to incorporate and compare multiple scale transformations in its modeling. The best exposure metric is selected based on a univariate model's *p*-value or deviance (ΔD) as specified. If a covariate search is specified in the control script, the final model is developed using backward elimination. *PoissonERM* identifies and avoids highly correlated covariates in the final model development of each endpoint. Predicting event incidence rates using external (simulated) exposure metric data is an additional functionality in *PoissonERM*, which is useful to understand the event occurrence associated with certain dose regimens. The summary outputs of the cleaned data, model developments, and predictions are saved in the working folder and can be compiled into a HTML report using Rmd.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Poisson regression has been used for count data, or binary data parameterized as an incidence rate, for many years, but there is no existing R package

Jessica Wojciechowski was an employee of Pfizer at the time of this work.

to conduct a comprehensive semi-automated exposure–response analysis via Poisson regression.

WHAT QUESTION DID THIS STUDY ADDRESS?

This tutorial focuses on implementing Poisson regression for exposure–response analysis of binary outcomes with an associated time and demonstrates how to conduct such analyses using the R package "*PoissonERM*". The statistical theory and assumptions when using Poisson regression are presented to ensure users of the R package understand exactly what is being conducted and can determine if it is appropriate for their data.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

This tutorial introduces the methodology and assumptions when using the Poisson regression model for incidence rates derived from binary endpoints and shows the usage of the new R package "*PoissonERM*". Detailed explanations regarding the methodology for the comprehensive exposure–response analysis performed by the package are provided, along with several examples demonstrating the modeling approach, results, simulations, and generated report.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

Poisson regression is a convenient option for exposure–response analysis of incidence rates derived from binary outcomes with an associated time component. The package not only simplifies the implementation of Poisson regression but also standardizes the modeling process for a complete and comprehensive exposure–response analysis.

INTRODUCTION

Exposure–response analysis for adverse event data

Adverse event (AE) data is usually collected in clinical studies for evaluating safety and reporting the seriousness and frequency associated with the use of a drug in patients. The AE information could be collected as part of the uniform collection for a registry (solicited), or it could be reported voluntarily (unsolicited).[1] Many AE data are collected as either binary outcomes with levels of "Yes" and "No" or multi-level categorical outcomes with several event types or several seriousness levels. When data is collected as categorical data with multiple levels, it is frequently dichotomized into binary outcomes for evaluating adverse events.

Exposure–response (ER) analysis is a widely used approach for understanding the relationship between a drug's exposure and efficacy or safety endpoints from a clinical study. In an exposure–response analysis on binary-response AE data, one challenge is the dependence on the duration of follow-up to observe events. A small number of observed events could be a result of a short duration rather than a low incidence rate, and it brings more complexity when data contains observations of different durations. On the other hand, low event frequencies may increase the uncertainties in the estimates. Therefore, it is important to conduct exposure–response analysis using appropriate models with binary data for valid interpretations and informative inferences.

Poisson regression for binary-outcome analysis

The Poisson distribution is a discrete probability distribution that describes count data. The distribution of the count data has a known constant mean with a variance equal to the mean. Poisson regression is a generalized linear model that can evaluate covariate effects as a linear function of the mean. The use of Poisson regression for the analysis of event counts is a common application within drug development, especially with regards to modeling the incidence rate of adverse events.[2–6]

Logistic regression is widely used for modeling binary outcomes; however, logistic regression might not be appropriate for "prospective and cross-sectional studies."[7] The odds ratio from logistic regression approximates the prevalence ratio well for rare events but it tends to overpredict for more common events, and the log-binominal regression model is preferred in such circumstances. The main disadvantages of log-binominal regression are that the standard error is underestimated (leading to a narrower confidence interval) and convergence problems. Poisson regression is a fully parametric approach to modeling the incidence rate of certain events, and it is a good alternative method for binary outcomes because it estimates the prevalence ratio correctly and its confidence intervals stay more conservative than with the log-binominal regression model.[7,8] Additionally, Poisson regression could handle the effect of time more properly because the Poisson model is a "memoryless" process (independence between events) and "homogeneous" (constant incidence rate within time intervals).[9] The performance of Poisson regression compared to logistic regression has been discussed in the literature.[10,11]

The scope of this tutorial is to showcase the implementation of Poisson regression with binary outcome data for an exposure–response analysis using an automated tool. There are some important assumptions for Poisson regression: (1) the distribution of the response is an integer-valued count per unit of time and follows a Poisson distribution where the mean and variance are equal, (2) the occurrence of the event is independent of the time since the last observed event (i.e., a memoryless process), and (3) the log of the mean rate must be a linear function of the covariates being evaluated.[9] The consequence of the second assumption is that the rate of the event is assumed to be fixed and constant over time (i.e., a homogenous process). Should the occurrence of the event make a participant more likely to experience the event again, then the second assumption would be violated and the inferences would be biased. To avoid these model assumptions, survival models and Cox proportional hazards models are other approaches that may be considered to model time-to-event,[5] while Markov models may be an option to model the transitions between states if such a dynamic nature is present and must be characterized.[9,12]

Automation tool: *PoissonERM*

To accelerate exposure–response analyses, automation tools have been developed to perform standard data summarization and model development with established methods.[13–15] The *PoissonERM* (Poisson Exposure–Response Modeling) R package was developed for the application of exposure–response modeling of binary event data that is commonly observed in clinical trials or other areas of drug development. All of the scripts used in the package provided with this tutorial have been reviewed for quality control (QC). The complete process of model development is automated within the package, beginning with data processing, covariate evaluation, model selection, model diagnostics, predictions, and finishing with an R markdown script that is generated, which can be knit into an HTML file summarizing all the results. There is also the ability to predict the incidence rates for new doses or exposures if the user provides the simulated exposures (usually from a separate population pharmacokinetic [PK] model). The complete workflow is shown in Figure 1. Automating the Poisson regression analysis not only accelerates the ER analysis but also reduces the burden for those who are new to Poisson regression.

The complete analysis is performed using the R statistical and programming language with the generalized linear model *glm()* function for estimation.[16] The analyses are conducted in the following manner:

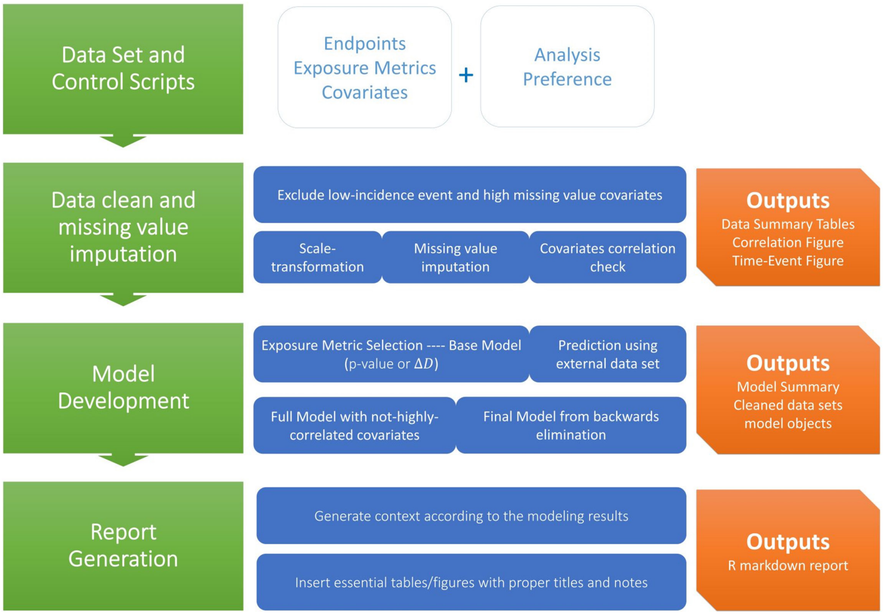
Data Set and Control Scripts
Endpoints
Exposure Metrics
Covariates
+
Analysis Preference
Data clean and missing value imputation
Exclude low-incidence event and high missing value covariates
Scale-transformation
Missing value imputation
Covariates correlation check
Outputs
Data Summary Tables
Correlation Figure
Time-Event Figure
Model Development
Exposure Metric Selection ---- Base Model (p-value or ΔD)
Prediction using external data set
Full Model with not-highly-correlated covariates
Final Model from backwards elimination
Outputs
Model Summary
Cleaned data sets
model objects
Report Generation
Generate context according to the modeling results
Insert essential tables/figures with proper titles and notes
Outputs
R markdown report

FIGURE 1 *PoissonERM* workflow.

- Univariate models are estimated to compare and select the best exposure metric, if any. The base model will only contain the selected exposure metric and an intercept. The base model selection is conducted based on the specified significant criterion using either the *p*-value or the deviance.
- Predicted cumulative incidence rates from the base model are created as figures and tables if requested by the user.
- If covariates are evaluated, then a full covariate model is estimated with the selected best exposure, if any, along with all covariates specified by the user.
- If covariates are to be evaluated, then a backward elimination procedure is run to determine the final model based on the specified significance criterion.
- The final model, including covariates, is then estimated.
- Risk ratio figures and tables are created if any demographic covariates are included in the final covariate model.
- If requested, predictions using an external exposure data set are calculated using the base model.

PoissonERM generates an R Markdown file that shows all the essential analysis components (summary tables, model details, conclusions, etc.) in an abbreviated report. The generated report explains the modeling results to users, and users could use this report as the initial draft of a comprehensive analysis report.

IMPLEMENTATION OF POISSON REGRESSION IN *POISSONERM*

Poisson distribution and the binomial distribution

Poisson regression of an event occurrence rate for the population in a given time interval is based on the asymptotic convergence of a binomial distribution with a low rate π and a large number of times the event is tested. When the count of the events follows a binomial distribution:

$$Y \sim \text{Binomial}(t, \pi) \tag{1}$$

where t is some unit of time and π is the constant rate of the event over time, then when t is large and π is small, the distribution can be approximated as

$$Y \sim \text{Poisson}(\lambda) \tag{2}$$

where $\lambda = t\pi$, and λ is the mean count of the event in the given time interval. The event rate can then be modeled using Poisson regression with the following linearization

$$\log(\lambda) = \log(t) + \log(\pi) \tag{3}$$

where $\log(t)$ is a time offset in the model.

PoissonERM is designed to conduct this type of analysis, where the offset for time is built into the linearization of the model. The full analysis can be conducted for multiple endpoints, with the models developed separately for each endpoint.

Poisson regression model development

A detailed chapter on the methodology of Poisson regression for binary endpoints can be found in the package repository.[17] Only the basic methodology and strategy for model development are presented herein.

Base model description

In the base model, the user-specified exposure metrics are tested as predictors for the occurrence of an event. The linear, square-root-transformed, and log-transformed scales of exposure metrics can be evaluated. Using a log-transformation is not recommended when the data contains placebo-treated subjects since their exposure will be 0, though the log-transformed exposure metric value log(0) is converted to log(0.0001) as a default in *PoissonERM* to avoid infinity values.

The counts of events are assumed to be characterizable using Poisson regression. The linear form of the regression is defined as a function of the mean (event rate). The base model includes an intercept and potential exposure metric as:

$$\log\left(\frac{\lambda}{t_j}\right) = \beta_0 + \beta_1 \cdot f\left(C_j\right) \tag{4}$$

which is equivalent to

$$\log(\lambda) = \beta_0 + \log\left(t_j\right) + \beta_1 \cdot f\left(C_j\right) \tag{5}$$

where $\lambda = E\left(Y_j\right)$ is the mean count during an time interval j, β_0 is the mean count when the time interval length $t_j = 1$ (unit), $\log\left(t_j\right)$ is the time-interval offset in log-scale, $f\left(C_j\right)$ is the appropriate exposure metric associated with the time interval j driving the response in mean count, and β_1 is the estimable effect of exposure on the mean count. More details regarding the unit of the time interval length t_j and the model prediction interpretation are provided in Section Prediction of incidence rates.

When assessing each exposure metric individually in the base model, the difference in the number of identifiable parameters (and therefore the degrees of freedom [df]) is equal to 1. Decision-making during model building is guided by an evaluation of the change in deviance, or $-2\cdot$ {log-likelihood}, between models. The deviance (DEV) is calculated as:

$$D = -2\log\left(\frac{L_0}{L_F}\right) \tag{6}$$

where L_0/L_F is the ratio of the likelihood of the null and fitted models. D can be shown to be approximately χ^2 distributed, with df equal to the difference in the number of parameters estimated between the null and fitted models.

The default criteria in the exposure metric selection is $p \leq 0.01$ where the exposure metric with the smallest p-value is selected, even if multiples meet the criteria. The user could provide a different significance level or select the ΔD-based criteria where the exposure metric with the largest change in deviance will be selected regardless of the significance.

Covariates assessment

PoissonERM conducts covariate assessment in 3 stages:

- The user specifies whether to consider transformations for all the continuous covariates. Only one of the original scales and the log-transformed scale will be selected for each continuous covariate before the model development. The selection is based on the p-value from a test of normality.
- Once the base model is determined for each endpoint, univariate models are estimated to select between highly correlated continuous covariates ($|r| \geq 0.6$, Spearman correlation coefficient) and between highly correlated categorical covariates ($|r| \geq 0.6$, Cramer's V). If any covariate needs to be tested based on physiological relevance, the user should avoid including covariates that are highly correlated with the required covariate.
- All selected covariates are added to the full model, and the covariate will be removed via stepwise backward elimination.

Full model development

To assess the E–R relationship between potential covariates and each of the endpoints, covariate effects are added to the linearization of the mean in Equation 4 as follows:

$$\log(\lambda) = \beta_0 + \log(t_j) + \beta_1 \cdot f(C_j) + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n \tag{7}$$

where λ is the arithmetic mean of the counts occurring during a certain time interval j, β_0 is the estimated intercept, β_1 is a regression coefficient (slope) representing the effect of the exposure metric, if any, determined in the base model development, and $\beta_2, \ldots, \beta_n$ represent the fixed effects, if any, of each additional covariate effect on the log-odds of the event occurring.

Final model development

The final model development starts with estimating the full model, which contains the parameters from the base model and the additional covariates under consideration. The full model is then subjected to a stepwise backward elimination procedure.

To compare two nested models, the difference in the deviance of each of the models follows an approximately χ^2 distribution with df equal to the difference in the number of estimated parameters:

$$\left(\frac{D_{\text{nested}} - D_{\text{full}}}{\text{df}_{\text{nested}} - \text{df}_{\text{full}}}\right) \sim \chi^2_{\text{df}_{\text{nested}} - \text{df}_{\text{full}}} \tag{8}$$

where $\text{df}_{\text{nested}}$ and df_{full} are the degrees of freedom for the nested and full models, respectively.

The default significance level for a covariate to remain is $p \leq 0.01$ (equivalent to $\Delta D \geq 6.63$). The user is able to specify a different significance level if desired. The default is that the exposure metric will not be evaluated in the backward elimination procedure, ensuring it is in the final model; however, the user may specify that it be tested together with all of the covariates.

Low incidence rate of an endpoint

The E–R analysis is performed for all endpoints where the number of events is higher than the threshold provided by the user. If the number of events is small, the maximum likelihood (ML) coefficients are known to be biased, and the uncertainty in the estimated incidence rate in Poisson regression will be underestimated due to the assumption that mean equals variance in Poisson distribution. It is recommended to only perform the analysis for the endpoint where the incidence rate is higher than 10%; otherwise, if the analysis is performed, the results should be interpreted with extra caution. In the case when an analysis is not performed for an endpoint, the event count will be shown in the demographic summary table in the abbreviated report, but the results section for that endpoint will not be created.

Handling of missing data

Missing data imputation is automatically performed within covariates, provided the percentage of missing values is less than or equal to the threshold percentage provided by the user. For continuous covariates, the missing values are imputed with the median value. Categorical covariates impute the mode. If the percentage of missing values is greater than the stated threshold, the covariates will automatically not be evaluated for the endpoint. These covariates are still summarized in demographic tables. Median imputation is implemented in the data cleaning step. If users would prefer a different imputation approach,[18] they should implement it in their input data set.

Nonsignificant exposure–response relationships

If the base model contains no exposure metric or the p-value of the selected exposure metric is higher than 0.05, the E–R relationship will be concluded as "not significant", regardless of the p-value significance level provided by the user. The predicted incidence rate curve and the confidence interval region will not be displayed in the prediction figure using external simulated exposure metrics.

Prediction of incidence rates

The predicted incidence rate in a given time interval is calculated using the base model:

$$\hat{\lambda} = \exp\left(\hat{\beta}_0 + \log(t) + \hat{\beta}_1 \cdot f(c)\right) \tag{9}$$

where $\hat{\lambda}$ is the predicted incidence rate in time interval t, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated intercept and coefficient, and $f(c)$ is the exposure metric associated with time point t. The predicted incidence rate per time unit $(t = 1)$ is $\exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot f(c)\right)$ for $\log(1) = 0$.

Additionally, the incidence rate is commonly reported as "incidence per 100 patient years":

$$\begin{aligned}&\text{Incidence per 100 patient years} = \\ &100 \times \exp\left(\hat{\beta}_0 + \log(1\text{year}) + \hat{\beta}_1 \cdot f\left(c_{1\text{year}}\right)\right)\end{aligned} \tag{10}$$

where $f\left(c_{1\text{year}}\right)$ is the exposure metric related to 1 year (365.25 days) of the treatment.

PoissonERM expects the time t to be provided in days and uses $t = 365.25$ (days) when predicting the incidence rate using external simulated exposure data. Then the incidence per 100 patient years is predicted using the base model as:

$$\begin{aligned}&\text{Incidence per 100 patient years} = \\ &100 \times \exp\left(\hat{\beta}_0 + \log(365.25) + \hat{\beta}_1 \cdot f\left(c_{365.25\text{ days}}\right)\right)\end{aligned} \tag{11}$$

If a different unit of time is used, the user could still calculate the predicted incidence rate by plugging in any time value in days and the exposure value at that time.

DATA STRUCTURE

PoissonERM expects the user to provide a vertical data structure with a flag column indicating the endpoint. If a simulated exposure data set is also provided, all data sets must be stored in the same directory.

Observation data for modeling

Table 1 shows the essential information of an analysis data set in order to be used for modeling using *PoissonERM*. Most of the column names are customizable except the "C" column, which is a row-exclusion indicator. An example of the observation data structure is provided in Table S1. For the data structure shown in Table S1, "PROT" can be used as the demographic-summary-by variable, "FLAG" and "SUB" are endpoint flag and sub-event type, respectively, "SEX" ~ "BWT" are the covariates, "CAVE1" and "CAVE2" are the exposure metric candidates, and "DV" is the observed outcome. Details for the levels of each categorical variable (such as sex, race, etc.) need to be provided in the control scripts, and the user should provide a formal name for each level.

External exposure metric data for prediction

When an external exposure data set (usually simulated from a PopPK model for certain regimens) is provided for predictions with *PoissonERM*, all the exposure metric values should be associated with 1 year (365.25 days) of treatment since the time for the prediction is fixed to be 1 year. The predicted event incidence rate (%) is then interpreted as the "incidence per 100 patient years" (see Section Prediction of incidence rates). The time information provided in the external data set will be ignored since the time is fixed at 1 year for predictions.

In order to generate predictions from the model, the data set must include:

TABLE 1 Columns in the observation data set for modeling.

Column	Inclusion	Column name	Details
Indicator of row exclusion	Mandatory	C	Any non-NA value indicates dropping the row, and an NA value indicates including the row in modeling
Subject ID	Mandatory	Customizable	The same ID indicates the same subject
Outcome	Mandatory	Customizable	0 indicates "not observed with event" and 1 indicates "observed with event." If there are multiple endpoints, all the outcomes are stored in this column
Endpoint flag	Mandatory	Customizable	The same value indicates the outcomes are for the same endpoints
Sub-event type in each endpoint	Optional	Customizable	Provide sub-categories of "Yes" and "No" for each endpoint. The value will be interpreted for each endpoint
Grade in each endpoint	Optional	Customizable	Provide the grade information for each endpoint, if the endpoint is grade-related
Time	Mandatory	Customizable	Time must be recorded in Days
Exposure metrics	Mandatory	Customizable	Each column is an exposure metric in original scale
Covariates	Optional	Customizable	Each column is a covariate. The values in categorical variables are recommended to be numbers regarding the level order
Demographic summary-by	Optional	Customizable	All covariates, endpoints, and exposure metrics will be summarized by this variable. It could be one of the provided covariates or a new variable

- A column containing the group label (usually the treatment regimen). The provided external exposure metrics and the predicted incidence rate will be summarized by this variable.
- Column(s) containing exposure metrics. It is recommended to provide the same exposure metrics as were provided in the observation data for modeling (all on the original scale). The minimal requirement is to provide all the exposure metrics that were selected for at least one endpoint.

An example of the external (simulated) exposure metric data structure is provided in Table S2. For the data structure shown in Table S2, "GROUP" is the group label, and "CAVE1" and "CAVE2" are the exposure metrics. The details of the variable names and the groups identified by the group label should be provided in the control script.

CONTROL SCRIPTS

The essential information for the analysis is provided in the control scripts – written in basic R code. Example scripts are provided in the supplemental material. The primary options in the modeling control script and the prediction control script that require the user's input are shown in Tables 2 and 3. For the modeling control script, most of these options have no default values since they are specific to the data set and the analysis. The user must provide the inputs to reflect the data set structure, exposure metrics, and analysis objectives. Apart from the main analysis options, the rest of the inputs for the modeling are all optional with default settings. Those options and their default settings are presented in Table S3. If the variable names are consistent between the analysis data set and the prediction data set, then only the file name and the grouping variables are required in the prediction user input.

USAGE

The complete statistical modeling, including data checks, data imputation, covariate screening, model selection, predictions, and model summaries, is all wrapped into the *ModelPoisson()* function. Only a control script and a data set are required. All of the output from the modeling is saved into summary folders and endpoint-specific folders.

PredictionPoisson() is the function generating additional prediction results, and *ReportPoisson()* compiles essential results into one automated R markdown report.

The package can be installed via GitHub,[17] where a vignette is available:

```
#Sys.setenv(R_REMOTES_STANDALONE = "true")
install.packages("devtools")
library(devtools)
```

TABLE 2 Key options requiring user's input in the modeling control scripts.

Value object	Class	Can be missing?	Default value	Details
input.data.name	Character	No	–	Data file for modeling. Must be under the working directory
pat.num	Character	No	–	Column of the subject ID in the data set
EVDUR	Character	No	–	Column of time (must be in days) in the data set
EVDUR.unit	Character	No	–	The unit of time. Recommended to use "days" or "Days"
dv	Character	No	–	Column of the binary outcome in the data set. It could contain the outcome of multiple endpoints. Usually contains 2 unique values. NA value is OK
dv.levels	Numeric or character	No	–	The unique values in dv column. Must be a vector of 2 values
dv.labels	Character	No	–	The labels of each level in the dv column. Must be a vector of 2 values
endpcolName	Character	No	–	Column of Event Flag. The data set may contain records of multiple endpoints; the same value in this column indicates the outcome for the same Event type
endpoints	Numeric or character		–	The Event Flags to be considered in this analysis. Must be a vector of at least one value. The value(s) must be included in endpcolName column
endpName	List	No	–	A list of each Event Flags and its corresponding name is shown in the report. Created via the sapply function. See the example script for more details
sub.endpcolName	Character	Yes	–	Column of Event Sub-type information. The unique value with the same Event Flag indicates the same sub-type (nested)
sub.endpName	List	Yes	–	A list of each Event Flags and its sub-types. See the example script for more details
dvg	Character	Yes	–	Column of Grade for all Event Flag
orig.exposureCov	Character	No	–	A vector of exposure metric columns
desc.exposureCov.1	List	No	–	A list of information for each exposure metric, generated via the sapply function. See the example script for more details
full.cat	Character	Yes only when full.cat1 is missing	–	A vector of categorical covariates
full.cat.1	List	Yes only when full.cat is missing	–	A list of information for each categorical covariate, generated via the sapply function. See the example script for more details
orig.con	Character	Yes only when orig.con.1 is missing	–	A vector of continuous covariates
orig.con.1	List	Yes only when orig.con is missing	–	A list of information for each continuous covariate, generated via the sapply function. See the example script for more details
demog_grp_var	Character	Yes	"PROT"	Column of summary-by variable in the data set. All events, exposure metrics, and covariates are summarized by this variable If the provided/default column does not exist in the data set, a column named "PROTnew" will be created to avoid running errors

```
install_github("yuchenw2015/PoissonERM",
build = FALSE)
```

More information about the package, including troubleshooting and package dependencies, is available in the package vignette.[17]

Statistical analysis: *ModelPoisson()*

```
rm(list = ls(all = TRUE))
folder.dir <- getwd()
ModelPoisson(pathRunType = folder.dir,
             user.input = “user-input.r”,
             clean = TRUE,
             save.name = “myEnvironment.
             RData”)
```

- `pathRunType`: The directory where the control scripts and data sets are. All the modeling results will be saved to this directory as well. This must be a fixed **absolute** path. The default value is *getwd()*.
- `user.input`: The file path of the control script (user-input.r) to source. Default value is NULL. If *user.input* is NULL, the user needs to run/source the control file first then run *ModelPoisson()* as

```
rm(list = ls(all = TRUE))
folder.dir <- getwd()
source(“user-input.r”)
ModelPoisson(pathRunType = folder.dir,
             user.input = NULL,
             clean = TRUE,
             save.name = “myEnvironment.
             RData”)
```

- `clean`: If *TRUE*, this will clean the folders under the directory *pathRunType* before running a new analysis. The default value of *clean* is *TRUE*.
- `save.name`: The modeling results and user's modeling options are saved as .RData file with the provided name under the directory *pathRunType*. The default value is “myEnvironment.RData”.

If the analysis is conducted successfully, several folders will have been created.

- “Demog-Sum” contains summary tables of events, exposures, and covariates.
- “Cov-EDA” contains figures of covariate correlation, exposure summary, and event summary vs. time.
- A folder is created for each endpoint. There will be subfolders within each of the endpoint folders containing:
  - an .RData file containing the post-processing data set used in modeling this endpoint;
  - “Models” folder containing the saved modeling R objects, the model summary tables, and the backward deletion log.
  - “OR” folder containing the saved Odds-Ratio tables and figures if the final model contains any covariates.

Predictions using external exposure metrics: *PredictionPoisson()*

PoissonERM provides an easy way to use the base model from *ModelPoisson()* to predict the incidence rate for new simulated exposures and compare it with the observed incidence rates grouped by exposure level.

If there is a significant exposure–response relationship (default $p < 0.05$ in the base model), *PredictionPoisson()* will generate the summary of observed incidence rates (in the data set used in *ModelPoisson()*), the distribution summary of the external exposure metric, and the predicted incidence rate with a 95% confidence interval for each group center (or the external exposure data center) calculated using the base model. If the exposure–response relationship is not significant, the prediction will not be calculated.

```
folder.dir <- getwd()
PredictionPoisson(pathRunType = folder.dir,
     prediction.input = “prediction-user-
     input-sim.R”,
     model.RData = “myEnvironment.RData”,
     save.name=“myEnvironment_new.RData”)
```

PredictionPoisson() creates one folder “Prediction” under each endpoint folder, which contains a new exposure summary and an incidence rate summary. The figures are saved in multiple sizes and can be used as needed.

- `pathRunType`: The directory where the control scripts and data sets are. All the results will be saved to this directory as well. It must be an absolute path. The default value is *getwd()*.
- `prediction.input`: The file path of the control script (prediction-user-input-sim.r) to source. The default value is NULL. If *prediction.input* is NULL, the user needs to run/source the control file first, then run *PredictionPoisson()* as

TABLE 3 Key options in the prediction control scripts.

Value object	Class	Can be missing?	Default value	Details
bin_n_obs	Numeric	Yes	7	Number of bins for grouping the exposure metric values in the observation data set
sim_inc_expo_data	Numeric	No	–	External exposure metric data file. All exposure metrics are associated with 1 year/52 weeks of treatment
Obs_Expo_list	Character	Yes	Provided column names of exposure metrics in the saved modeling results	Column name of exposure metrics in the observation data set. It could be a subset of all exposure metrics as long as the metrics selected for each endpoint are included
Sim_Expo_list	Character	Yes	Same value as Obs_Expo_list	Column name of exposure metrics in the external exposure metric data set. The names correspond to names in Obs_Expo_list. It is recommended to use the same column name for the same original exposure metric in the observation data set and the external exposure metric data set
Center_Metric	Character	Yes	"geomean"	The geomean of the exposure metric in each group is reported by default. The other choice is "median"
Center_Metric_name	Character	Yes	Determined based on Center_Metric	Name of the center metric to show in figures and tables
grp_colname	Character	Yes	–	Column of group label in the external exposure metric data set. If missing, the exposure metric and the prediction will be summarized as a whole group
grp_colname_tab	Character	Yes	"Label"	Name of the group label to show in tables and figures
levels.grp	Numeric or character	Yes	Unique values in group label column	A vector of levels in the group label column
labels.grp	Numeric or character	Yes	Same value as levels.grp	A vector of names for each level in the group label column
filter_condition	Character	Yes	–	A string of the condition to subset the data set
expo_pred_tab_caption	Character	Yes	"Predicted exposure metric for each dose are derived from simulated patients with randomly drawn random effect parameters as described by the final population PK model and body weights sampled from observations."	Explanations for the simulated exposure metrics are shown in the footnotes of each figure

```
folder.dir <- getwd()
source("prediction-user-input-sim.R")
PredictionPoisson (pathRunType = folder.
dir,
    prediction.input = NULL,
    model.RData = "myEnvironment.Rdata",
    save.name = "myEnvironment_new.Rdata")
```

- `model.Rdata`: The saved modeling result object from *ModelPoisson()* must be located under the directory *pathRunType*. The default value is "myEnvironment.Rdata", which is the default *save.name* value in the function *ModelPoisson()*.
- `save.name`: The previous modeling results and the prediction results are saved as. Rdata with the provided name under the directory *pathRunType*. The default value is *model.Rdata*, which will overwrite the previously saved modeling result.

The automated predictions with *PoissonERM* are made using the base model. The base, full, and final models of each endpoint are saved as R objects should the user wish to conduct any further analyses (including predictions) manually.

Report generation: *ReportPoisson()*

The modeling results from *ModelPoisson()* or *PredictionPoisson()* can be used to generate an automated abbreviated report via R Markdown rendered as HTML. The user should run *ReportPoisson()* right after running *ModelPoisson()* to use the results from *ModelPoisson()* in the report, or run *ReportPoisson()* right after running *PredictionPoisson()* to use the results from *PredictionPoisson()* in the report.

If *PredictionPoisson()* has already been performed but the user does not want to include the predictions in the report, user may remove those "Prediction" folders manually or simply clean the R Environment and rerun *ModelPoisson()* with *clean = TRUE*.

```
folder.dir <- getwd()
ReportPoisson(pathRunType = folder.dir,
      model.RData = "myEnvironment.RData",
      file.name = "Report.Rmd")
```

- `pathRunType`: The directory where the previously saved modeling results (folders and the saved .RData file) are located. The generated .Rmd file will be saved to this directory. It must be an absolute path. The default value is *getwd()*.
- `model.RData`: The saved modeling result object from *ModelPoisson()* or from *PredictionPoisson()*, must be located under directory *pathRunType*. The default value is "myEnvironment.RData," which is the default save.name value in the function *ModelPoisson()*.
- `file.name`: The file name for the generated report. The default value is "NULL." If *file.name* is NULL or if *file.name* is not a string ended with ".Rmd," the default file name will be "Poisson-Regression-Date&Time.Rmd."

EXAMPLE ANALYSIS USING *POISSONERM*

A data set was generated to demonstrate the usage of *PoissonERM*. In this simulated data set, there are three endpoints ("Adverse Event 1," "Adverse Event 2," "Adverse Event 3"), five demographic covariates, and two simulated exposure metrics ("C_{ave_1}" and "C_{ave_2}"). The control scripts and the generated report are included in the supplemental materials, while the full analysis is available on Github.[19]

Using the simulated data set and the example control scripts, the following code runs the full ER analysis on the three endpoints.

```
library(PoissonERM)
rm(list = ls(all = TRUE))
folder.dir1   <-   "PoissonERM/Example1/"
#change the absolute path accordingly
ModelPoisson(pathRunType = folder.dir1,
    user.input = "user-input.r")
PredictionPoisson(pathRunType = getwd(),
    prediction.input = "prediction-user-
    input-sim.R",
    model.RData = "myEnvironment.RData")

ReportPoisson(pathRunType = getwd(),
      model.RData="myEnvironment.RData",
      file.name = "Report_with_pred.Rmd")
```

Data summary and covariates/endpoint exclusion

PoissonERM reported the proportion of each category for all categorical covariates and the median, range, mean, and standard deviation (SD) for all continuous covariates (Table S4). In this example, all variables were summarized by protocol ("PROT"), which is the default "summary-by" variable. The user can assign different covariates to the model development of the different endpoints, and the user could provide "summary-only" covariates that are only used in the summary table. In this example of a simulated data set, all covariates provided were included in the model development of each endpoint and there are no "summary-only" covariates.

The proportion of missing values was reported for each covariate in the summary table (Table S4). In this example, there were no missing values in any of the covariates, so there was neither imputation nor covariate exclusion. The correlation between covariates was examined before the model development. In this example, there were neither highly correlated categorical nor continuous covariate pairs (Figure S1).

Figure 2 shows the exposure summary for one of the endpoints (Adverse Event 2). The boxplot shows that the median C_{ave_2} for participants observed with the event was higher than the median C_{ave_2} for those with no events, while there was no obvious difference in the median C_{ave_1} between the participants with and without observed events. A few of the log-transformed C_{ave_1} and C_{ave_2} points are located far away from the bulk because these C_{ave_1} and C_{ave_2} values were 0 (placebo) and the log-transformation converted them to log(0.0001). It is not recommended to consider log-transformation when there are placebo arms in the data.

Table 4 shows the event counts for one of the endpoints (Adverse Event 2). There were 4 and 2 subcategories in "Yes" and "No". The user could choose to show the summary of sub-categories in the event by providing the information about sub-categories in the control script. The endpoint "Adverse Event 1" was ignored in the analysis because the incidence rate of 5.5% (Table S5) was lower than the provided threshold of 10%. The event time was summarized for each endpoint (Figure S2). The events were observed at different time points, and the histograms of the event counts suggested a uniform distribution; that is, the event occurrence rate was similar for each period of time, which implies that the assumption of "constant rate over time" was not violated.

Model development

The exposure metric selection criteria was specified as "*p*-value," and the significance level was 0.01. The selected exposure metric for Adverse Event 2 was C_{ave_2} (Table S6) since it had the smallest *p*-value (also the largest change in deviance) among all 6 candidate exposure metrics (including the transformed ones). The base model (Table S7) was used to predict the incidence rate for Adverse Event 2 using external simulated exposure data from hypothetical dose regimens (Figure 3). All the simulated exposure metrics correspond to 1 year of treatment (or steady-state exposures). The exposure–response relationship was not significant in Adverse Event 3; consequently, the prediction figure only shows the observed incidence rate and the distribution of the external exposure data (Figure S3, Adverse Event 3).

The backward elimination was conducted with a full model containing all five covariates with a significance level of 0.01 (Figure S4). It was specified that the selected exposure metric was to be evaluated as part of the

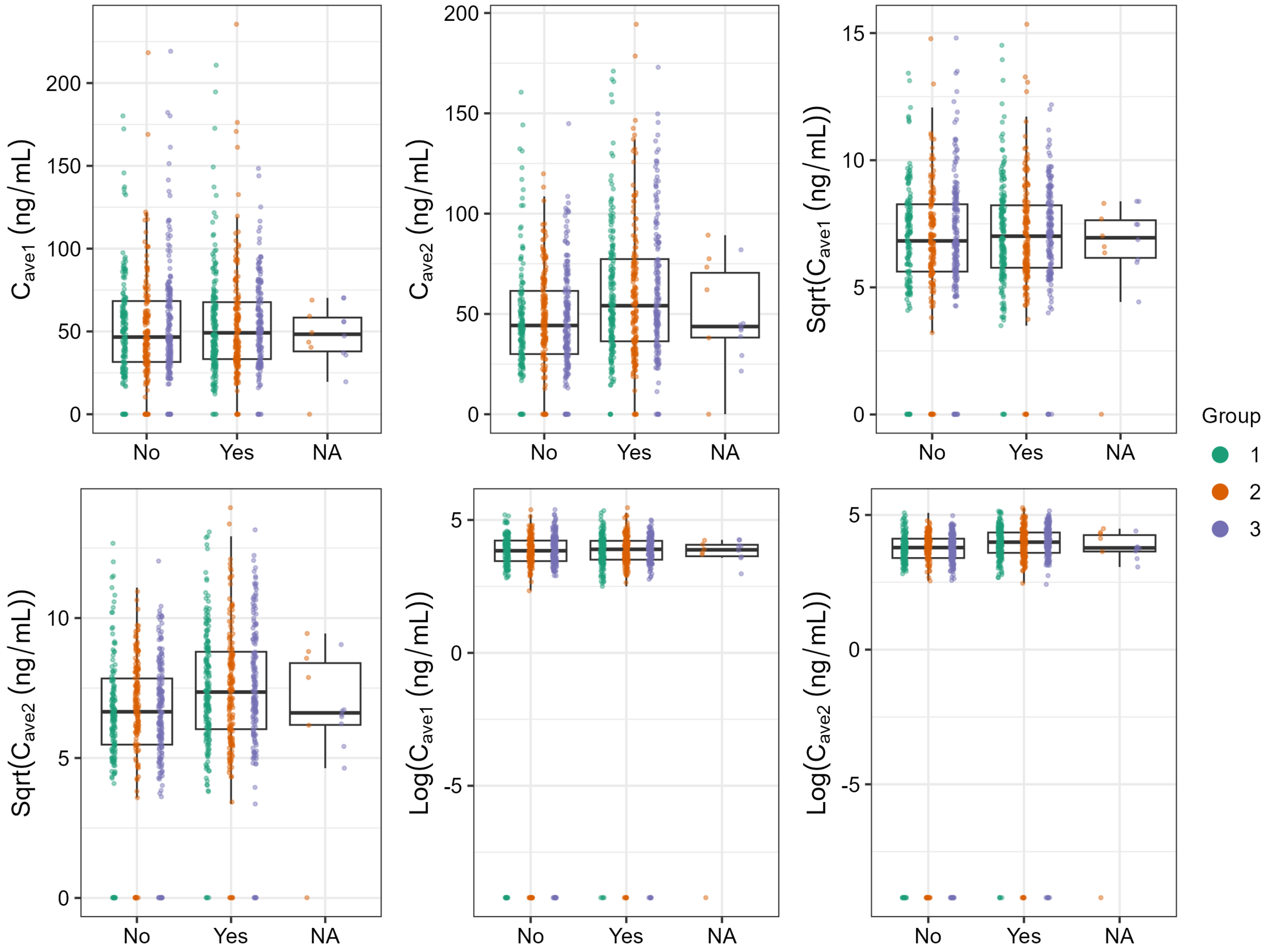

FIGURE 2 An example of the exposure metrics summaries. Exposures were grouped by the dependent variable as "No" (0, no event observed), "Yes" (1, event observed), and "NA" (missing value), and the specified group-by variable "Protocol" as "1," "2," and "3."

backward elimination procedure. For endpoint Adverse Event 2, "Geographical Location" was identified in the final model (Table 5); for endpoint Adverse Event 3, "Sex" was identified in the final model, though there was no exposure metric in the model. Additional plots were generated to show the odds ratio of each identified covariate with a 95% confidence interval (Figure S5).

TABLE 4 Example of event counts (Automated Table).

Statistics	1	2	3	Total
Total number of individuals	327	327	346	1000
Adverse event type 2				
Event = Yes				
SubType 1	45	45	37	
SubType 2	48	39	44	
SubType 3	48	43	29	
SubType 4	36	39	51	
Event = No				
SubType 5	74	79	90	
SubType 6	76	76	87	
Missing (%)	0 (0)	6 (1.8)	8 (2.3)	14 (1.4)

Note: Percentage values refer to the number of patients.

Analysis conclusions

Brief conclusions for each analyzed endpoint were automatically generated. Adverse Event 1 was ignored in the model development due to the low incidence rate; thus, no conclusion was drawn for Adverse Event 1. The ER analysis for Adverse Event 2 and Adverse Event 3 was auto-summarized as follows:

A Poisson regression model was developed for Adverse Event Type 2.

- ○ The selected best exposure metric was C(ave2) (ng/mL) ($p < 0.0001$).
- ○ For the model of Adverse Event Type 2, the analysis data set included 1000 study participants. 504 of them reported events, and 482 of them reported non-events.
- ○ There was one covariate in the final model: Geographical Location.
- ○ The covariate Geographical Location, with the reference level of Geographical Location US, resulted in a statistically significant improvement in the model fit. The odds ratios and 95% confidence intervals for Geographical Location Non-US are 1.96 (1.64, 2.34), respectively.

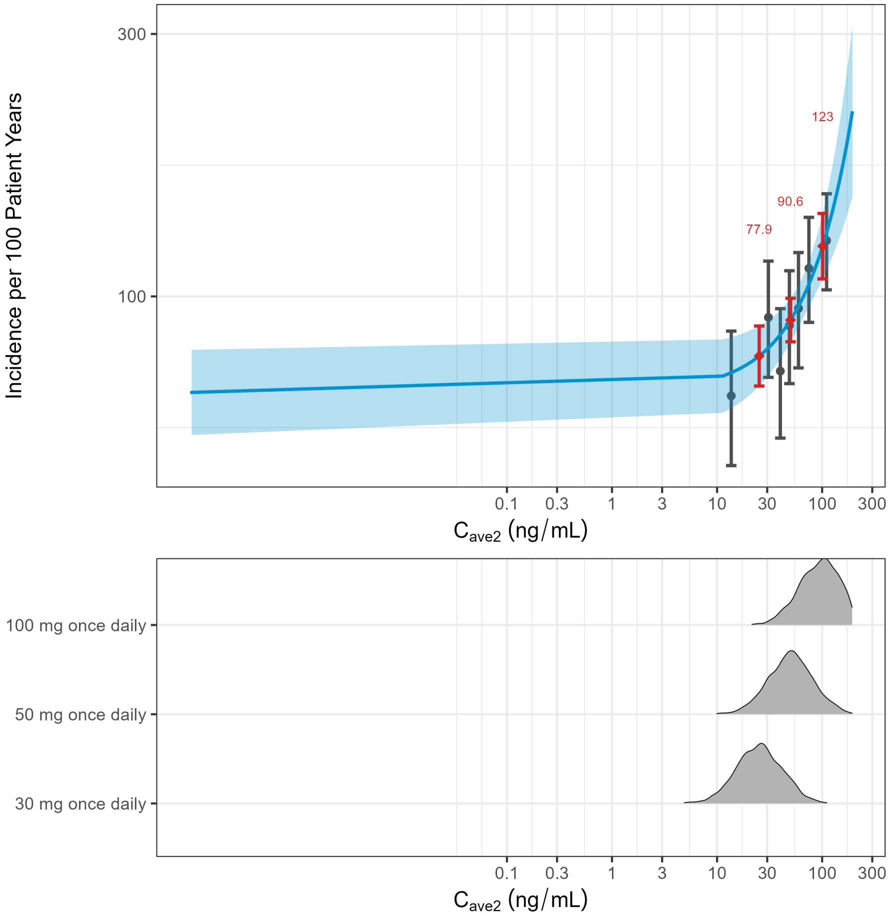

FIGURE 3 An example of predictions using the base model for Adverse Event 2. Top: Gray circles and error bars are the observed mean and 95% CI, respectively, of the incidence of Adverse Event Type 2 (all types) per 100 patient years for each bin of the shown exposure metric in the analysis population (n of bins = 7). The blue line (and blue shaded area) are the model-predicted mean (and 95% CI) incidence per 100 patient years for the range of observed exposure metrics in the analysis population. Red circles and error bars are the model-predicted mean and 95% CI, respectively, for the incidence per 100 patient years. Bottom: Predicted exposure values for each dose are derived from 2500 simulated subjects.

TABLE 5 Example of estimates from the final model (Automated Table).

Effects	Estimate	95% CI	*z* value	Pr(>|*z*|)	Odds ratio	95% CI of odds ratio
(Intercept)	−6.54	(−6.73, −6.34)	−66.7	<0.001		
Geographical location Non-US	0.673	(0.496, 0.848)	7.49	<0.001	1.96	(1.64, 2.34)
C(ave2) (ng/mL)	0.00595	(0.00336, 0.00847)	4.56	<0.001	1.01	(1, 1.01)
n	***N***	**Δ*D***	**df**	**Pr(>Chisq)**	**LogLik**	**AIC**
986	1000	73.44	2	<0.001	−880.34	1766.68

Note: Confidence Intervals calculated based on the log-likelihood function (profile confidence interval). *n* = Number of participants with event outcomes recorded. *N* = Total number of participants in the study.

A Poisson regression model was developed for Adverse Event Type 3.

- There was not a statistically significant exposure-response relationship.
- For the model of Adverse Event Type 3, the analysis data set included 1000 study participants. 312 of them reported events, and 681 of them reported non-events.
- There was one covariate in the final model: Sex.
- The covariate Sex, with the reference level of Sex Male, resulted in a statistically significant improvement in the model fit. The odds ratios and 95% confidence intervals for Sex Female are 1.66 (1.32, 2.09), respectively.

DISCUSSION

Exposure–response analyses are key components of a comprehensive assessment of the safety of new investigational therapies. It is typical for event-based safety outcomes to be reported as the incidence rate across the population or over a period of time. Poisson regression is a useful option for modeling incidence rates to evaluate potential ER relationships when it is important to also account for time rather than just the occurrence of the event. Provided the assumptions are considered reasonable, Poisson regression can provide fast results due to its simplicity of implementation.

It is important that users evaluate whether the assumptions inherent in Poisson regression are reasonable for their specific data, particularly the assumption that the event is "memoryless." If any of the assumptions are violated, other classes of models should be considered (e.g., survival models, cox models, Markov models, etc.). There is no formal statistical test to check the "memoryless" assumption; therefore, researchers should use their own judgment to decide if this assumption is reasonable. Researchers could examine the observation data to check if there is any state-related pattern in the occurrence of events. *PoissonERM* automates the creation of figures for the event counts against time, which can reflect the population-level incidence rate over time. Other models, such as non-parametric hazard rate estimation, can be applied to examine the assumption if the data support it, but such models are usually more complex than Poisson regression analysis.

There are several choices for the exposure metric for the ER analysis. Time-weighted exposure (TWE) is commonly used for capturing temporal dynamics and reflecting cumulative exposure, while it can introduce confounding effects in the ER model as TWE is affected by multiple factors[20] and the interpretation can be challenging. Simpler metrics like maximum concentration or steady-state concentration will simplify the model interpretation, but such metrics might not be appropriate for events that occurred before achieving the steady state, titration schemes, dose reductions, or studies where the duration was not long enough. Another option is to calculate the average concentration over a certain time (such as a dosing cycle or the time between two adjunct visits), which is not event-time-related but still reflects the fluctuation in exposure. As the exposure is an input that must be calculated prior to using *PoissonERM*, the user must use their own judgment when selecting appropriate exposure metric candidates.

PoissonERM was developed to automate the complete exposure–response analysis of binary outcomes with an associated time and generate an abbreviated report. As a fully quality-control-reviewed package, *PoissonERM* could be used as a semi-automated analysis tool to conduct a comprehensive analysis as well as an auxiliary tool to help users start with Poisson regression analysis as an exploratory assessment. The results are fully reproducible when provided with the analysis data set, the prediction data set (when applicable), and the control scripts. The generated R markdown file can then be used to share preliminary results or as an advanced starting point to write a formal report. By automating the entire analysis, *PoissonERM* provides users with a quick and easily implementable option to consider Poisson regression when it is appropriate for their data.

AUTHOR CONTRIBUTIONS

Y.W., L.F., J.W., D.I. and T.N wrote the manuscript; Y.W., L.F., J.W., D.I. and T.N designed the research; Y.W and L.F. performed the research; Y.W. and J.W. analyzed the data.

FUNDING INFORMATION

This research was funded by Pfizer Inc.

CONFLICT OF INTEREST STATEMENT

Yuchen Wang is a current employee of Pfizer and may own stock in Pfizer. Luke Fostvedt is a current employee of Pfizer and may own stock in Pfizer. Jessica Wojciechowski is a former employee of Pfizer and may own stock in Pfizer. Donald Irby is a current employee of Pfizer and may own stock in Pfizer. Timothy Nicholas is a current employee of Pfizer and may own stock in Pfizer.

ORCID

Yuchen Wang https://orcid.org/0009-0000-2351-2090
Luke Fostvedt https://orcid.org/0000-0002-6714-1188
Jessica Wojciechowski https://orcid.org/0000-0002-3302-8742
Timothy Nicholas https://orcid.org/0000-0001-8863-665X

REFERENCES

1. Gliklich RE, Dreyer NA, Leavy MB. Adverse event detection, processing, and reporting. *Registries for Evaluating Patient Outcomes: A User's Guide [Internet]*. 3rd ed. Agency for Healthcare Research and Quality (US); 2014.
2. Muralidharan KK, Steiner D, Amarante D, et al. Exposure-disease response analysis of natalizumab in subjects with multiple sclerosis. *J Pharmacokinet Pharmacodyn*. 2017;44:263-275. doi:10.1007/s10928-017-9514-4
3. Combes FP, Baneyx G, Coello N, et al. Population pharmacokinetics-pharmacodynamics of oral everolimus in patients with seizures associated with tuberous sclerosis complex. *J Pharmacokinet Pharmacodyn*. 2018;45:707-719. doi:10.1007/s10928-018-9600-2
4. Jonsson F, Schmitt C, Petry C, Mercier F, Frey N, Retout S. Exposure-bleeding count modeling of emicizumab for the prophylaxis of bleeding in persons with hemophilia A with/without inhibitors against factor VIII. *Clin Pharmacokinet*. 2021;60:931-941. doi:10.1007/s40262-021-01006-0
5. Yang W, Jepson C, Xie D, et al. Statistical methods for recurrent event analysis in cohort studies of CKD. *Clin J Am Soc Nephrol*. 2017;12:2066-2073. doi:10.2215/CJN.12841216
6. Frome EL, Checkoway H. Use of Poisson regression models in estimating incidence rates and ratios. *Am J Epidemiol*. 1985;121:309-323. doi:10.1093/oxfordjournals.aje.a114001
7. Fekedulegn D, Andrew M, Violanti J, Hartley T, Charles L, Burchfiel C. Comparison of statistical approaches to evaluate factors associated with metabolic syndrome. *J Clin Hypertens*. 2010;12:365-373. doi:10.1111/j.1751-7176.2010.00264.x
8. McNutt L-A, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*. 2003;157:940-943. doi:10.1093/aje/kwg074
9. Plan E. Modeling and simulation of count data. *CPT Pharmacometrics Syst Pharmacol*. 2014;3:10-12. doi:10.1038/psp.2014.27
10. Cleophas TJ, Zwinderman AH. Poisson regression for binary outcomes (52 Patients). *SPSS for Starters and 2nd Levelers*. Springer; 2016:273-277. doi:10.1007/978-3-319-20600-4_47
11. Ijomah M, Biu E, Mgbeahurike C. Assessing logistic and poisson regression model in analyzing count data. *International Journal of Applied Science and Mathematical Theory*. 2018;4:42-68.
12. Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. *Med Decis Mak*. 1993;13:322-338. doi:10.1177/0272989x9301300409
13. Irby D, Fostvedt L, Nickens D, Wang Y, Nicholas T. ERMod: a semi-automated Exposure–Response (E–R) analysis and reporting tool to support decision making while increasing time and cost savings in drug development (poster). American Conference of Pharmacometrics 13; 2022 November; Aurora, CO, USA.
14. Wojciechowski J. Utility of automation and pro-active modeling. American Conference on Pharmacometrics 13; 2022 November; Aurora, CO, USA.
15. Wang Y, Fostvedt L, Irby D, Wojciechowski J, Huh Y. ERMod poisson: a semi-automated exposure–response (E–R) analysis and reporting tool with prediction feature (poster). American Conference of Pharmacometrics 14; 2023 November; Oxon Hill, MD, USA.
16. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2013.
17. Wang Y, Fostvedt L. R package: PoissonERM. 2024 https://github.com/yuchenw2015/PoissonERM
18. Johansson ÅM, Karlsson MO. Comparison of methods for handling missing covariate data. *AAPS J*. 2013;15:1232-1241. doi:10.1208/s12248-013-9526-y
19. Wang Y, Fostvedt L. Examples for PoissonERM. 2024 https://github.com/yuchenw2015/PoissonERM-Example
20. Wiens MR, French JL, Rogers JA. Confounded exposure metrics. *CPT Pharmacometrics Syst Pharmacol*. 2024;13:187-191. doi:10.1002/psp4.13074

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wang Y, Fostvedt L, Wojciechowski J, Irby D, Nicholas T. Automated Poisson regression exposure–response analysis for binary outcomes with *PoissonERM*. *CPT Pharmacometrics Syst Pharmacol*. 2024;13:1615-1629. doi:10.1002/psp4.13207