Secondary Publication

Appelbaum, Sebastian; Stronski, Julia; Konerding, Uwe; u. a.

The Use of Double Poisson Regression for Count Data in Health and Life Science : A Narrative Review

Date of secondary publication: 09.10.2025

Version of Record (Published Version), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-110660x

Primary publication

Appelbaum, Sebastian; Stronski, Julia; Konerding, Uwe; u. a. (2025): The Use of Double Poisson Regression for Count Data in Health and Life Science : A Narrative Review, in: Stats : open access journal of statistical sciences, Basel : MDPI, Vol. 8, Nr. 4, 90, pp. 1–12, doi: 10.3390/stats8040090.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

stats

MDPI

Review

The Use of Double Poisson Regression for Count Data in Health and Life Science—A Narrative Review

Sebastian Appelbaum [1,2], Julia Stronski [1], Uwe Konerding [1,3] and Thomas Ostermann [1,*]

[1] Department of Psychology and Psychotherapy, Witten/Herdecke University, Alfred-Herrhausen-Str. 50, 58448 Witten, Germany; Sebastian.appelbaum@uni-wh.de (S.A.); julia.stronski@uni-wh.de (J.S.); uwe.konerding@uni-bamberg.de (U.K.)
[2] Center for Clinical Trials, Department of Medicine, Faculty of Health, Witten/Herdecke University, 58448 Witten, Germany
[3] Trimberg Research Academy, Otto-Friedrich-Universität Bamberg, An der Weberei 5, 96047 Bamberg, Germany
* Correspondence: thomas.ostermann@uni-wh.de

Abstract

Count data are present in many areas of everyday life. Unfortunately, such data are often characterized by over- and under-dispersion. In 1986, Efron introduced the Double Poisson distribution to account for this problem. The aim of this work is to examine the application of this distribution in regression analyses performed in health-related literature by means of a narrative review. The databases Science Direct, PBSC, Pubmed PsycInfo, PsycArticles, CINAHL and Google Scholar were searched for applications. Two independent reviewers extracted data on Double Poisson Regression Models and their applications in the health and life sciences. From a total of 1644 hits, 84 articles were pre-selected and after full-text screening, 13 articles remained. All these articles were published after 2011 and most of them targeted epidemiological research. Both over- and under-dispersion was present and most of the papers used the generalized additive models for location, scale, and shape (GAMLSS) framework. In summary, this narrative review shows that the first steps in applying Efron's idea of double exponential families for empirical count data have already been successfully taken in a variety of fields in the health and life sciences. Approaches to ease their application in clinical research should be encouraged.

Keywords: count data; Double Poisson; regression; generalized additive models

check for updates

Academic Editors: Bogdan Oancea, Adrian Pană and Cătălina Liliana Andrei

Received: 20 August 2025
Revised: 24 September 2025
Accepted: 26 September 2025
Published: 1 October 2025

Citation: Appelbaum, S.; Stronski, J.; Konerding, U.; Ostermann, T. The Use of Double Poisson Regression for Count Data in Health and Life Science—A Narrative Review. *Stats* **2025**, *8*, 90. https://doi.org/10.3390/stats8040090

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Count data are present in many areas of everyday life. In interpersonal relationships, count data may include the number of children in a family [1,2] or the number of marriages or divorces [2,3]. In sports, the number of goals scored in soccer [4] or the number of targets missed in biathlon [5] often determine success or failure of teams and athletes. In passenger transportation, count data appear in accident statistics [6] and sick days are an important parameter for health insurance companies when calculating costs [7]. Moreover, as early as the 19th century, the number of cancer cases was documented in cancer registers as an important marker for the health of the population [8]. To put it briefly, count data appear in many different areas of everyday life, epidemiology and health care.

All examples described have one thing in common: count data are characterized by non-negative integers. In many cases, they exhibit a right skew in the distribution, i.e., lower count values are more common than higher values. As a result, established

methods of inferential statistics are only applicable to a limited extent. For this reason, very early, models were sought that are better suited for analyzing count data [9]. The first contribution to this research was the Poisson distribution introduced by Siméon Denis Poisson in 1837 [10]. Due to its mathematical properties, this distribution still plays a dominant role in the analyses of count data [11].

The Poisson distribution is based on an exponential approach to modeling the discrete probability distribution of events occurring in a fixed interval of time or space. If the mean event rate μ is known, then the probability of a certain number of y events follows the density function given by

$$f_{\mu}(y) = \frac{\mu^y}{y!} e^{-\mu} \tag{1}$$

In the case of regression analysis of count data, the generalized linear model with Poisson distributed outcome thus can be written as

$$ln(E[Y|x]) = \beta_0 + \beta' x \tag{2}$$

where $x \in \mathbb{R}^n$ is the vector of predictor variables and $\beta \in \mathbb{R}^n$ is the vector of the regression coefficients.

The exponential values of these regression coefficients represent the Incidence Rate Ratio (IRR). The IRR is a statistical measure that quantifies the multiplicative change in the expected rate of events for a one-unit increase in the predictor variable, holding all other predictors constant. With this relationship, the regression coefficients can be interpreted meaningfully in terms of a measure of effect: while an IRR of 1 denotes no change in the incidence rate, IRRs greater or lower than one indicate a higher or lower incidence rate under exposure. If, for example, the IRR for a particular predictor is 1.7, this means that for every one-unit increase in the predictor, the incidence rate of the event increases by 70%.

An unpleasant property of the Poisson distribution in (1) is that its mean is equal to its variance, which is usually referred to as equi-dispersion. Therefore, in the case of over- and under-dispersion, i.e., in cases when the observed variance is higher or, respectively, lower than the mean, Poisson regression leads to incorrect estimates of the standard errors and thus to incorrect test statistics [12]. The problem of insufficient modeling of under-dispersion also applies to negative binomial distribution, introduced by Yule in 1910 [13] and Greenwood and Yule in 1920 for the analysis of count data [14]:

$$f_{\mu,\,\alpha}(y) = \frac{\Gamma\left(\alpha^{-1} + y\right)}{\Gamma(\alpha^{-1})\Gamma(y+1)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu}\right)^{y} \tag{3}$$

To account for this problem, Consul and Jain in 1973 [15] introduced a two-parameter discrete probability distribution which allows the variance to be greater than, equal to, or less than the mean [16]:

$$f_{\mu,\,\delta}(y) = \frac{\mu(\mu + \delta y)^{y-1} e^{-\mu - \delta y}}{y!} \tag{4}$$

To some extent, this approach permits modelling the case of under-dispersion. Unfortunately, in this approach, the parameter δ can take high negative values [17] in some cases of under-dispersion. In such cases, the density function is not defined and therefore truncated to zero. For this reason, this approach is also referred to as the Restricted Generalized Poisson regression model [18].

Another approach named after its inventors is the so-called Conway–Maxwell distribution [19], i.e.,

$$f_{\mu,\,\nu}(y) = \frac{\mu^y}{(y!)^{\nu} Z(\mu, \nu)} \tag{5}$$

where $Z(\mu, \nu) = \sum_{j=0}^{\infty} \frac{\mu^j}{(j!)^\nu}$ is a normalizing constant. This distribution permits to handle all kinds of dispersion [20]. However, if the original form of this distribution is applied in a regression model, the resulting regression coefficients are difficult to interpret. Sellers and Shmueli in 2010 showed that the relationship between the conditional mean and the predictors in the Conway–Maxwell–Poisson regression approach is neither additive nor multiplicative [21]. Although this has since been solved with a mean-parameterized Conway–Maxwell–Poisson regression [22], this problem may have deterred researchers from using this approach at that time. (see Table 1 for an overview).

Table 1. Distributions for count data.

Distribution	Dispersion [a]			Expected Value E(Y) [b]	Variance [b]	Interpretation of Coefficients [c]
	Under-	Equi-	Over-			
Poisson	−	+	−	$= \mu$	$= \mathrm{E}(Y)$	IRR
Negative Binomial	−	+	+	$= \mu$	$= \mathrm{E}(Y) + \alpha \mathrm{E}(Y)^2$	IRR
Restricted Generalized Poisson	(+)	+	+	$= \frac{\mu}{1-\delta}$	$= \frac{1}{(1-\delta)^2} \mathrm{E}(Y)$	IRR
Conway–Maxwell–Poisson	+	+	+	$\approx \mu^{1/\nu} - \frac{\nu-1}{2\nu}$	$\approx \frac{1}{\nu} \mathrm{E}(Y)$	(IRR) after mean-parameterization
Double Poisson	+	+	+	$\approx \mu$	$\approx \frac{\mathrm{E}(Y)}{\alpha}$	IRR

[a] +: can be modeled; −: cannot be modeled; (+): can be modeled with restrictions. [b] μ is the location parameter; α, δ, ν are dispersion parameters taken from the Formulas (1) to (5) and (7). [c] IRR: Incidence Rate Ratio.

In the time before the reparameterization of the Conway-Maxwell-Poisson distribution was known, Efron [23] tried to find a distribution that was flexible enough to be applied to count data showing over-, equi-, or under-dispersion and whose results were easy to interpret in terms of content. For this purpose, he used an extension of the one-parametrical exponential families model. He applied the double exponential family with the density function

$$f_{\mu,\alpha,n}(Y = y) = \alpha^{1/2} \{g_{\mu,n}(y)\}^{\alpha} \{g_{y,n}(y)\}^{1-\alpha} [dG_n(y)] \tag{6}$$

and a dispersion parameter α to families $g_{\mu,n}(y)$ which originally had only one unknown parameter μ. Substituting the density of the Poisson distribution in $g_{\mu,n}(y)$ leads to

$$f_{\mu,\alpha}(y) = \alpha^{1/2} e^{-\alpha\mu} \left(\frac{e^{-y} y^y}{y!}\right) \left(\frac{e\mu}{y}\right)^{y\alpha}, \quad \alpha > 0 \tag{7}$$

As stated above, μ_i is defined as

$$\mu_i = \exp\left(\beta_0 + x_i^T \beta\right) \tag{8}$$

in the case of a log-linear regression and, thus, the formula for Double Poisson Regression is

$$f(y_i|\mu_i, \alpha) = c(\mu_i, \alpha)\ \alpha^{1/2} e^{-\alpha\mu_i} \left(\frac{e^{-y_i} y_i^{y_i}}{y_i!}\right) \left(\frac{e\mu_i}{y_i}\right)^{y_i\ \alpha}, \quad \alpha > 0 \tag{9}$$

with α being a parameter that models dispersion, and $c(\mu_i,\alpha)$ being a normalization factor:

$$c(\mu_i, \alpha) = \frac{1}{\sum_{y=0}^{\infty} \alpha^{1/2} e^{-\alpha \mu_i} \left(\frac{e^{-y} y^y}{y!}\right) \left(\frac{e\mu_i}{y}\right)^{y\alpha}} \cong \frac{1}{1 + \frac{1-\alpha}{12\alpha\mu_i}\left(1 + \frac{1}{\alpha\mu_i}\right)} \quad (10)$$

According to [23], the approximation given in the right-hand term of this formula is usually close to one, which guarantees that the sum of all probability mass function values equals one. Thus, there are two strategies to handle $c(\mu_i,\alpha)$: setting it equal to one as proposed in [24] or using the approximation proposed in [23] in the right term of the formula. Efron in [23] also showed that the conditional means and corresponding conditioned variances of the predictor values and the dispersion parameter are

$$E[Y|\mu_i, \alpha] \cong \mu_i \text{ and } Var[Y|\mu_i, \alpha] \cong \frac{\mu_i}{\alpha} \quad (11)$$

Accordingly, over-dispersion corresponds to a parameter $\alpha < 1$, while under-dispersion is described by $\alpha > 1$.

As illustrated above, the approach introduced by Efron for modeling over- and under-dispersion is mathematically quite complex, in particular with regard to the normalization factor $c(\mu_i, \alpha)$ [25], as has also been demonstrated in [26]. This might be the reason that it is not extensively available in conventional statistical software packages. On the other hand, it offers an optimal framework to account for over- and under-dispersion in count data. This gives rise to the question as to how extensively Double Poisson Regression has hitherto been applied in research. This question is addressed here by means of a narrative review. As mentioned at the outset, count data play a significant role in epidemiology, so this review will focus on work in the field of epidemiology and clinical research.

2. Materials and Methods

Although this review does not evaluate clinical outcome parameters, it follows the practice guide for conducting narrative reviews [27].

2.1. Inclusion and Exclusion Criteria

To be included in the review, the article had to be published in a journal, conference proceeding or a book and include an application of the Double Poisson Regression Model proposed by Efron in the field of the health and life sciences. Purely mathematical treatments of the Double Poisson Regression were excluded. As the term Double Poisson is also used for bivariate Poisson distributions, applications not dealing with the approach by Efron were also excluded. Finally, animal or plant research studies were excluded.

2.2. Search Strategy

The literature search was conducted in Science Direct, Google Scholar, and the EBSCOhost search and retrieval system [28] for the databases PubMed (MEDLNE), PsycInfo, PsycArticles, CINAHL, and Psychology and Behavioral Sciences Collection (PBSC). In an initial step, the search query ("Double Poisson" AND regression) was used. No further specifications (e.g., "study" or "health") were made to initially obtain a large pool of literature. In addition, articles citing Efron's original work were also screened for suitability.

2.3. Selection Strategy

The selection of relevant articles followed a 2-step process (see Figure 1). First, two reviewers initially screened the bibliography resulting from the search by examining titles and abstracts. In the second step, the full texts of the remaining articles were analyzed for

their final suitability. In cases of discrepancies between the reviewers, a third reviewer was consulted, and a majority principle was applied.

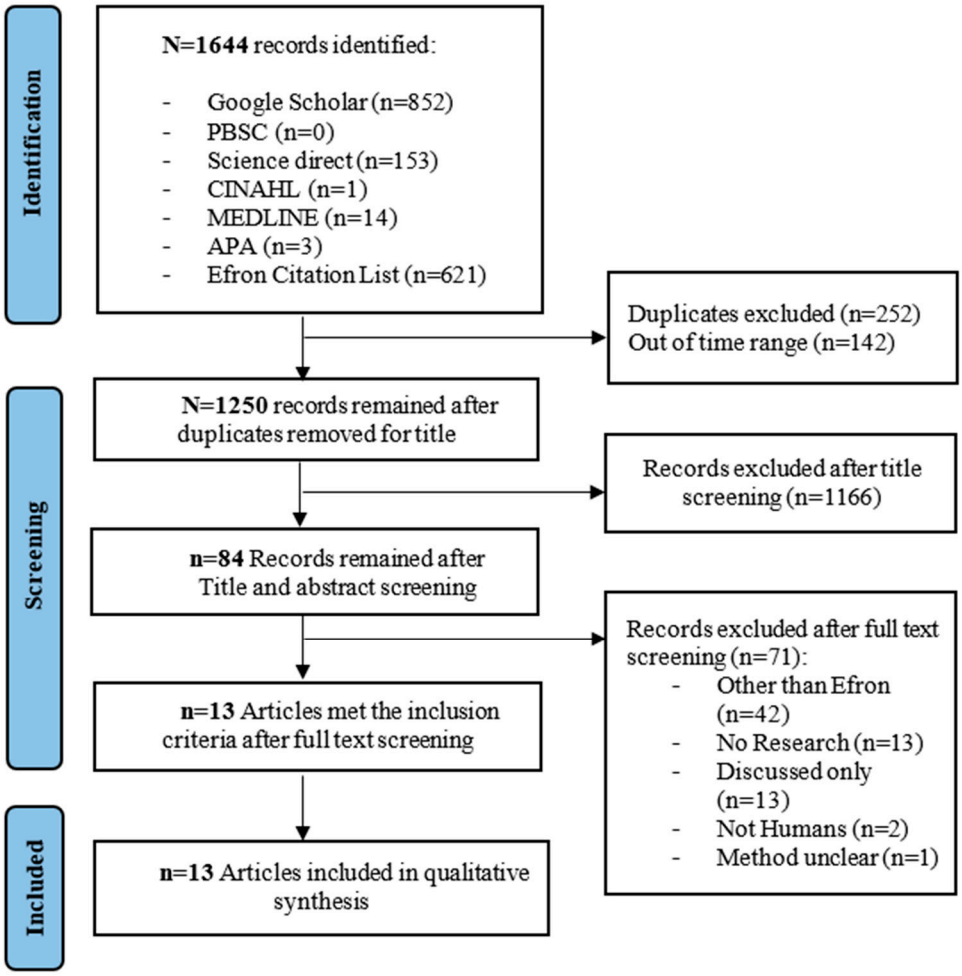

Figure 1. Overview and results of applying the search and selection strategy.

2.4. Data Extraction

The data of each article were imported into a single spreadsheet on Microsoft Excel. The data included the following parameters: Authors, Year, Area of research, Type of study, Type of count variable, Outcome presentation, Quality criteria applied (i.e., performance measures for comparing alternative specifications of regression models), Type of dispersion, Implementation, and Software used.

3. Results

With this search strategy, a total of 1644 articles were identified, most of which were found in the Google Scholar Search (n = 852) and Efron's citation list (n = 621), while Medline, for example, returned only 14 hits. After cleaning for duplicates and articles out of the time range (n = 394), title and (when possible) abstract screening of the remaining articles led to a pre-selection of 84 articles. After applying full text screening, 13 articles remained for data extraction. The whole process of applying the search and selection strategy is shown in Figure 1, while Table 2 gives an overview of the studies included.

Although the original publication by Efron dates back to 1986, it took a quarter of a century for the first publication by Gijbels & Prosdocimi [29] to apply this method to analyze data on the induced abortion rate in Italy provided by the Italian National Statistics Institute. As a special feature, the authors introduced P-splines to estimate the smooth components of

a proposed Generalized Additive Model (GAM). Moreover, they also used a new statistical approach to model the dispersion as a function of covariates and thus combined the GAM with the Double Exponential Family framework proposed by Efron [30].

Almost at the same time, Quintero-Sarmiento et al. published an application of Double Poisson Regression to infant mortality data in 36 areas for the 5-year period 2000–2005 in Colombia [31]. They found clear evidence for overdispersion in the data and applied different statistical models including covariates. They investigated fit to data using AIC and BIC and finally decided against Double Poisson Regression (AIC = 324.0; BIC = 338.3) and used a negative binomial model (AIC = 317.1; BIC = 329.8). As a conclusion, they recommended to initially determine the causes of the overdispersion in the data to be able to model them appropriately.

In 2016, a series of articles on Double Poisson Regression from Brazil started with the analysis of 14,419 snakebites in the state of São Paulo between 2007 and 2014 [32]. They used the Double Poisson Regression model for a Bayesian time series model of the snakebites on monthly bases, including intercept, linear trend, four parameters for a seasonal effect, and one parameter for an autoregressive term of first order to detect seasonal as well as regional predictors for different types of snakes. They recommended using Bayesian methods as a reasonable alternative to frequentist methods due to the complexity of the likelihood function.

Two further studies were subsequently launched from the same group of statisticians: based on an observational study of 41 patients suffering from recurrent respiratory papillomatosis, Nogueira et al. in [33] used a Double Poisson Regression approach to detect the impact of different sociomedical variables, such as age, on the number of surgeries in the patients. Unfortunately, no more information on quality criteria was given in this paper. In another article by Nunes et al. [34] the Double Poisson approach was only mentioned in the abstract and not in the statistical part or in the results.

At the same time as the first paper by Aragon et al. [32] was published, Phiri et al. presented a case study of Schistosomiasis and soil-transmitted helminth infections in Chikwawa, Malawi [35]. In addition to the Double Poisson approach, the bivariate Poisson, diagonal-inflated bivariate Poisson, and bivariate zero-inflated Poisson models were also considered to model egg counts of the parasites. As a result, they found that children aged 6–15 years with a low socio-economic status and a low level of education had an increased risk of co-occurring infections. Statistically, the diagonal-inflated bivariate Poisson model performed best with the lowest AIC and BIC values.

With the first comprehensive introduction of generalized additive models for location, scale, and shape (GAMLSS) in R in 2017 [36] Double Poisson Regression gained considerable popularity. The first application was presented in the paper by de Andrate in 2021 [37], in which they merged the geographical information of the 218 urban census districts and the 420 case counts of tuberculosis in the period from 2013 to 2018. To identify the best fitting model for the case counts, all families of distributions in the GAMLSS package were tested and evaluated using AIC values. In contrast to [35], Double Poisson Regression showed the lowest AIC values and thus was selected. As a result, income, gender, and age between 15 and 59 years were found to be significant predictors for cases of tuberculosis. Dispersion was modeled via distributional regression; however, the intercept was not reported.

Two years later, the same working group published a similar paper on tuberculosis. A total of 1730 individuals diagnosed with TB between 2001 and 2017 in 811 urban census districts within Macapá were included. Again, a Double Poisson Regression Model with seven covariates and two intercepts for μ and σ was chosen and an inclusion of quadratic effects was tested [38].

Hu et al. [39] considered all discrete distributions provided by the GAMLSS package. Based on a survey of 2760 adolescents between 16 and 19 years, the Adolescent Self-rating Life Events Checklist (ASLEC) was chosen as the count variable for the various regression models. The distribution with the smallest value of the Schwarz Bayesian criterion [40] was then chosen for further modeling. In contrast to [35] this was the Double Poisson distribution. Here, the quantitative covariates were smoothed using a local maximum-likelihood-penalized B-splines method.

GAMLSS was also used by Rajalu et al. [41], who modeled the daily counts of patients with traumatic brain injuries (TBI; n = 8893) between December 2019 and January 2021 by means of a bootstrapped time series approach with a Double Poisson distribution. To compare the profile of TBI cases seen before and during the COVID-pandemic, a subset of these cases, seen between 1 December 2019 and 31 July 2020 (n = 5259), are studied in detail. The model used by Rajalu et al. included four covariates for the expected value and three covariates to model the variance. Plotting the AIC over the observed time period, they found a change point in TBI cases at the beginning of the COVID-19-pandemic (20 March 2020) and a significant relationship between the number of TBI cases per day and alcohol sales on weekdays.

In a study of oral health, the Double Poisson Regression Model in GAMLSS was used to perform an age-adjusted regression analysis including eight covariates in 1158 study participants [42]. The authors found an association between obesity, socioeconomic status and oral health. The data showed over-dispersion, and effects were reported as ratios (exp(β)) of the regression coefficients. Other quality criteria were not reported.

Two epidemiological studies complete the picture of applications. In 2021, Khoei et al. investigated the number of congenital malformations reported in 6368 neonatal health records from Khoy, a city in Iran, in 2017 [43]. As a specialty, they found extremely high numbers of cases (99.6%) without congenital malformations. Consequently, they applied the Bayesian zero-inflated and Hurdle models [44]. Although the deviance information criterion (DIC) values for all applied models were quite similar, they found a Hurdle Double Poisson Regression model with two binary covariates to perform best in fitting the data with parental consanguinity as a prognostic factor (exp (0.59) = 1.80).

Orooji et al. applied a similar model in 2022 to determine prognostic factors for the number of coronary-artery-stenoses [45]. The data basis was derived from 633 elderly cardiovascular patients at Ghaem Hospital, Mashhad, Iran from September 2011 to May 2013. Since the number of coronary-artery-stenoses has a high zero inflation and cannot exceed three, a right truncated zero-inflated Double Poisson Regression model with seven covariates in the count part of the model and four in the logit part was applied. This model outperformed the non-truncated zero-inflated Double Poisson model with respect to a considerably lower AIC (1522.4 versus 1734.8) and a recognizable over-dispersion (σ = 4.3478). They found that Body Mass Index and female gender were significantly associated with the count part of the model.

Summary

In summary, applications of the Double Poisson Regression Model were mainly from the field of ecological or epidemiological studies, while clinical research was clearly in the minority of applications. Data were mainly extracted from official statistics or registers. Almost all studies used software packages like the R-package GAMLSS [46,47] or OpenBugs [48] which also reflects the increase in publications since the publication of relevant publications. To select the best fitting model, established quality criteria such as AIC or BIC were also reported in almost all publications found in this review. Surprisingly, this does not apply to the interpretation of regression coefficients as effect sizes. None of the studies

found here took advantage of the possibility of interpreting regression coefficients as IRRs, which has been proposed and published for other models [17,49].

Table 2. Articles included.

Author (Year), Country [a]	Area	Type of Study	Count Variable	Quality Criteria [b]	Type of Dispersion	Software
Aragon et al. (2016) [32], Brazil	Public Health	Ecological study	Snakebites	DIC	Under-dispersion	Open BUGS
de Andrade et al. (2021) [37], Brasil	Infectious diseases	Ecological study	Infectious diseases	AIC	Modeled Dispersion [c]	R/ GAMLSS
Giacomet et al. (2023) [38], Brazil	Infectious diseases	Ecological study	Infectious diseases	AIC	Over-dispersion	R/ GAMLSS
Gijbels & Prosdocimi (2011) [29], Italy	Obstetrics	Epidemiological study	Abortion Rate	AISE	Under-dispersion	R/ GAMLSS
Hu et al. (2023) [39], China	Sleep quality	Survey	PSQI	BIC	Modeled Dispersion [c]	R/ GAMLSS
Khoei et al. (2021) [43], Iran	Neonatology	Epidemiological study	Congenital malformations	DIC	Over-dispersion	Open BUGS
Nogueira et al. (2021) [33], Brazil	Diseases of the respiratory system	Observational study	Number of surgeries	Not reported	Over-dispersion	SAS
Nunes et al. (2021) [34], Brazil	Angioedema	Observational study	Angioedema attacks	Not reported	Not reported	Open BUGS
Orooji et al. (2022) [45], Iran	Cardio-logy	Epidemiological study	Number of vessels with stenosis	AIC	Under-dispersion	SAS
Phiri et al. (2016) [35], Malawi	Infectious diseases	RCT-Secondary analysis [d]	Co-occurrence of parasites	AIC and BIC	Over-dispersion	STATA
Quintero-Sarmiento et al. (2012) [31], Colombia	Mortality	Epidemiological study	Children < 5 years who died	AIC and BIC	Over-dispersion	Not reported
Rajalu et al. (2022) [41], India	Emergency medicine	Epidemiological study	Cases of traumatic brain injury	AIC	Over-dispersion	R/ GAMLSS
Schmidt et al. (2022) [42], Germany	Oral Health	Cohort study	Oral Health Score	Not reported	Over-dispersion	R/ GAMLSS

[a] Sorted alphabetically by first author. [b] AIC: Akaike information criterion; BIC: Bayesian information criterion; DIC: Deviance information criterion; AISE: Approximate integrated squared error. [c] "Modeled Dispersion" indicates that dispersion is modeled as a function of covariates. [d] RCT: Randomized controlled trial.

4. Discussion

This narrative review aims to present a body of literature of applications of the Double Poisson Regression Model in health care. Developed by Efron in 1986, it took 25 years to be used from 2011 onwards. Popularity of the model increased with the launch of the R-package GAMLSS in 2008 [46] and subsequent publications [36,47]. Due to the pre-set options for using different distributions in this package, there are now much simpler ways of applying these very complex models including the Double Poisson Regression Model. Further studies should clarify the extent to which biased estimators and model non-convergence can occur here and how results are comparable to other software, e.g., OpenBUGS [48].

Most of the applications were located in the field of epidemiology. As also stated by Hohberg et al. [50], the application of the Double Poisson Regression Model in the evaluation of RCTs is far away from being established and "beyond the mean". This is perhaps also the reason why there is still a clear heterogeneity in the presentation of the results and the application of quality indicators such as AIC or BIC, which in the case of [51] led to that study not being included in the list of results, as no specific results on Double Poisson Regression were presented. A corresponding checklist, such as that used in structural equation modeling [52], would certainly improve the reporting quality here.

Another important finding is that the Double Poisson Regression Model was used to model both over- and under-dispersed data, which is a clear strength of this approach. However, as could also be shown, the Double Poisson Regression is only one way of dealing with over- and under-dispersion [53], although a very elegant one from a mathematical perspective. Other approaches already mentioned here are negative binomial distribution [54] or double gamma distribution [55], which were also found within this review but excluded for this topic. In particular, in the case of under-dispersion, the other options provide either very limited possibilities or parameters that cannot be interpreted in a meaningful way [56].

Probably inspired by Efron's idea, other, e.g., three-parametric, methods have also been developed to model over-dispersion. For example, Nandi et al. in [57] describe a convolution of a Poisson variable and an independently distributed negative binomial random variable. However, use cases for this approach have not yet been published.

4.1. Limitations

This work has some important limitations. Unlike reviews in the field of clinical research, which often address a clearly defined question, a review of the application of a statistical method is much more complex. It is quite likely that studies dealing with Double Poisson Regression have gone undetected, even though we have attempted to minimize this bias, for example, by analyzing citations of Efron's original work.

Another limitation concerns the language bias: the search focused on English-language publications, which potentially missed relevant non-English studies. This limitation is exacerbated by a potentially restricted database selection. Although it is quite comprehensive, some regional or specialized databases containing gray literature may have been omitted.

Finally, there is a lack of formal quality assessment tools for this kind of study, which limits the ability to judge the robustness of the primary studies.

4.2. Further Research

Further research related to Double Poisson Regression should clarify how to deal with the normalization factor in practical applications. Initial simulations show clear dependencies in convergence behavior [26]. This may necessitate further developments in the software for Double Poisson Regression.

With regard to answering the question of when each model should be applied, the following research questions arise: (1) To what extent do deviations from the model assumptions from the data actually affect the estimation of the regression parameters and their standard errors or the quality of the inferential statistical decisions based on them? (2) Which statistics are best suited to capture deviations from the model assumptions that are relevant for regression? (3) How are these statistics related to the estimation of the regression parameters and their standard errors, or the quality of the inferential statistical decisions based on them? Answers to these questions could form the basis of an algorithm that can be used to decide, based on the data, when each regression model is best to use.

5. Conclusions

Statistical methods have always been adapted to empirical data. This also applies to the analysis of count data. One of the most mathematically sophisticated approaches is the family of double distributions introduced by Efron.

This narrative review shows that the first steps in applying this idea for empirical count data have already been successfully taken in a variety of fields in the health and life sciences. This seems to have been fostered by the availability of relevant software. Approaches to improve the application of double distributions in clinical research should be encouraged.

Author Contributions: S.A. and T.O.: conception of the work; J.S. and T.O.: data acquisition and selection; TO and SA: data analysis and interpretation; T.O., S.A., U.K., and J.S.: writing the manuscript; T.O. and U.K.: substantial revising of the manuscript. All authors approved the manuscript in the submitted version and take responsibility for the scientific integrity of the work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ratna, M.B.; Khan, H.A.; Hossain, M.A. Modeling the number of children ever born in a household in Bangladesh using generalized Poisson regression. *Ulab J. Sci. Eng.* **2011**, *3*, 51–56.
2. Winkelmann, R.; Zimmerman, K.F. Count data models for demographic data. *Math. Popul. Stud.* **1994**, *4*, 205–221. [CrossRef]
3. Ghaznavi, C.; Kawashima, T.; Tanoue, Y.; Yoneoka, D.; Makiyama, K.; Sakamoto, H.; Ueda, P.; Akifumi, E.; Nomura, S. Changes in marriage, divorce and births during the COVID-19 pandemic in Japan. *BMJ Glob. Health* **2022**, *7*, e007866. [CrossRef]
4. Loukas, K.; Karapiperis, D.; Feretzakis, G.; Verykios, V.S. Predicting football match results using a Poisson regression model. *Appl. Sci.* **2024**, *14*, 7230. [CrossRef]
5. Maier, T.; Meister, D.; Trösch, S.; Wehrlin, J.P. Predicting biathlon shooting performance using machine learning. *J. Sports Sci.* **2018**, *36*, 2333–2339. [CrossRef]
6. Pińskwar, I.; Choryński, A.; Graczyk, D. Good weather for a ride (or not?): How weather conditions impact road accidents—A case study from Wielkopolska (Poland). *Int. J. Biometeorol.* **2024**, *68*, 317–331. [CrossRef]
7. Denis, T.; Lanfranchi, J. A new empirical model of the determinants of sickness and the choice between presenteeism and absence. *Labour* **2025**, *39*, 61–87. [CrossRef]
8. Ostermann, T.; Appelbaum, S.; Baumgartner, S.; Rist, L.; Krüerke, D. Using merged cancer registry data for survival analysis in patients treated with integrative oncology: Conceptual framework and first results of a feasibility study. In Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies, Online, 9–11 February 2022; Volume 5, pp. 463–468.
9. Ondrick, C.W.; Griffiths, J.C. Fortran IV computer program for fitting observed count data to discrete distribution models of binomial, Poisson and negative binomial. In *Kansas Geological Survey Computer Contribution 35*; Merriam, D.F., Ed.; University of Kansas: Lawrence, KS, USA, 1969.
10. Poisson, S.D. *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile: Précédées des Règles Générales du Calcul des Probabilités*; Bachelier: Paris, France, 1837.
11. Coxe, S.; West, S.G.; Aiken, L.S. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *J. Pers. Assess.* **2009**, *91*, 121–136. [CrossRef]
12. Palmer, A.; Losilla, J.M.; Vives, J.; Jiménez, R. Overdispersion in the Poisson regression model. *Methodology* **2007**, *3*, 89–99. [CrossRef]
13. Yule, G.U. On the distribution of deaths with age when the causes of death act cumulatively, and similar frequency distributions. *J. R. Stat. Soc.* **1910**, *73*, 26–38. [CrossRef]
14. Greenwood, M.; Yule, G.U. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or repeated accidents. *J. R. Stat. Soc.* **1920**, *83*, 255–279. [CrossRef]
15. Consul, P.C.; Jain, G.C. A generalization of the Poisson distribution. *Technometrics* **1973**, *15*, 791–799. [CrossRef]
16. Consul, P.C.; Famoye, F. Maximum likelihood estimation for the generalized Poisson distribution when sample mean is larger than sample variance. *Commun. Stat.–Theory Methods* **1988**, *17*, 219–234. [CrossRef]

17. Harris, T.; Yang, Z.; Hardin, J.W. Modeling underdispersed count data with generalized Poisson regression. *Stata J.* **2012**, *12*, 736–747. [CrossRef]
18. Famoye, F. Restricted generalized Poisson regression model. *Commun. Stat.—Theory Methods* **1993**, *22*, 1335–1354. [CrossRef]
19. Conway, R.W.; Maxwell, W.L. A queuing model with state dependent service rates. *J. Ind. Eng.* **1962**, *12*, 132–136.
20. Shmueli, G.; Minka, T.P.; Kadane, J.B.; Borle, S.; Boatwright, P.B. A useful discrete distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **2005**, *54*, 127–142. [CrossRef]
21. Sellers, K.F.; Shmueli, G. A flexible regression model for count data. *Ann. Appl. Stat.* **2010**, *4*, 943–961. [CrossRef]
22. Huang, A. Mean-Parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Stat Model.* **2017**, *17*, 359–380. [CrossRef]
23. Efron, B. Double exponential families and their use in generalized linear regression. *J. Am. Stat. Assoc.* **1986**, *81*, 709–721. [CrossRef]
24. Aragon, D.C.; Achcar, J.A.; Martinez, E.Z. Maximum likelihood and Bayesian estimators for the double Poisson distribution. *J. Stat. Theory Pract.* **2018**, *12*, 886–911. [CrossRef]
25. Zou, Y.; Geedipally, S.R.; Lord, D. Evaluating the double Poisson generalized linear model. *Accid. Anal. Prev.* **2013**, *59*, 497–504. [CrossRef]
26. Appelbaum, S.; Ostermann, T.; Konerding, U. Maximum likelihood estimation of parameters for double poisson regression: A simulation study. *Comput. Stat.* **2025**, 1–39. [CrossRef]
27. Siddaway, A.P.; Wood, A.M.; Hedges, L.V. How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annu. Rev. Psychol.* **2019**, *70*, 747–770. [CrossRef]
28. Matthews, J.R. EBSCOhost. *Libr. Technol. Rep.* **1996**, *32*, 221–227.
29. Gijbels, I.; Prosdocimi, I. Smooth estimation of mean and dispersion function in extended generalized additive models with application to Italian induced abortion data. *J. Appl. Stat.* **2011**, *38*, 2391–2411. [CrossRef]
30. Gijbels, I.; Prosdocimi, I.; Claeskens, G. Nonparametric estimation of mean and dispersion functions in extended generalized linear models. *Test* **2010**, *19*, 580–608. [CrossRef]
31. Quintero-Sarmiento, A.; Cepeda-Cuervo, E.; Núñez-Antón, V. Estimating infant mortality in Colombia: Some overdispersion modelling approaches. *J. Appl. Stat.* **2012**, *39*, 1011–1036. [CrossRef]
32. Aragon, D.C.; Queiroz, J.A.M.D.; Martinez, E.Z. Incidence of snakebites from 2007 to 2014 in the State of São Paulo, Southeast Brazil, using a Bayesian time series model. *Rev. Soc. Bras. Med. Trop.* **2016**, *49*, 515–519. [CrossRef]
33. Nogueira, R.L.; Küpper, D.S.; do Bonfim, C.M.; Aragon, D.C.; Damico, T.A.; Miura, C.S.; Passos, I.M.; Nogueira, M.L.; Rahal, P.; Valera, F.C. HPV genotype is a prognosticator for recurrence of respiratory papillomatosis in children. *Clin. Otolaryngol.* **2021**, *46*, 181–188. [CrossRef]
34. Nunes, F.L.; Ferriani, M.P.; Moreno, A.S.; Langer, S.S.; Maia, L.S.; Ferraro, M.F.; Sarti, W.; de Bessa Junior, J.; Cunha, D.; Suffritti, C.; et al. Decreasing attacks and improving quality of life through a systematic management program for patients with hereditary angioedema. *Int. Arch. Allergy Immunol.* **2021**, *182*, 697–708. [CrossRef]
35. Phiri, B.B.; Ngwira, B.; Kazembe, L.N. Analysing risk factors of co-occurrence of schistosomiasis haematobium and hookworm using bivariate regression models: Case study of Chikwawa, Malawi. *Parasite Epidemiol. Control* **2016**, *1*, 149–158.
36. Rigby, R.A.; Stasinopoulos, M.D.; Heller, G.Z.; De Bastiani, F. *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*; CRC Press Taylor & Francis Group: Boca Raton, FL, USA, 2017.
37. De Andrade, H.L.P.; Arroyo, L.H.; Yamamura, M.; Ramos, A.C.V.; de Almeida Crispim, J.; Berra, T.Z.; Santos Neto, M.; Carvalho Pinto, I.; Palha, P.F.; Monroe, A.A.; et al. Social inequalities associated with the onset of tuberculosis in disease-prone territories in a city from northeastern Brazil. *J. Infect. Dev. Ctries.* **2021**, *15*, 1443–1452. [CrossRef]
38. Giacomet, C.L.; Ramos, A.C.V.; Moura, H.S.D.; Berra, T.Z.; Alves, Y.M.; Delpino, F.M.; Farley, J.E.; Reynolds, N.R.; Bodini Alonso, J.; Teibo, T.K.A.; et al. A distributional regression approach to modeling the impact of structural and intermediary social determinants on communities burdened by tuberculosis in Eastern Amazonia–Brazil. *Arch. Public Health* **2023**, *81*, 135. [CrossRef]
39. Hu, Y.; Duan, X.; Zhang, Z.; Lu, C.; Zhang, Y. Effects of adverse events and 12-week group step aerobics on sleep quality in Chinese adolescents. *Children* **2023**, *10*, 1253. [CrossRef]
40. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
41. Rajalu, B.M.; Devi, B.I.; Shukla, D.P.; Shukla, L.; Jayan, M.; Prasad, K.; Jayarajan, D.; Kandasamy, A.; Murthy, P. Traumatic brain injury during COVID-19 pandemic—Time-series analysis of a natural experiment. *BMJ Open* **2022**, *12*, e052639. [CrossRef]
42. Schmidt, J.; Vogel, M.; Poulain, T.; Kiess, W.; Hirsch, C.; Ziebolz, D.; Haak, R. Association of oral health conditions in adolescents with social factors and obesity. *Int. J. Environ. Res. Public Health* **2022**, *19*, 2905.
43. Khoei, R.A.A.; Kazemnejad, A.; Eskandari, F.; Heidarzadeh, M. Analysis of infant congenital malformation data using the Bayesian count regression. *Iran. Red Crescent Med. J.* **2021**, *23*, e180.
44. Green, J.A. Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. *Health Psychol. Behav. Med.* **2021**, *9*, 436–455.

45. Orooji, A.; Sahranavard, T.; Shakeri, M.T.; Tajfard, M.; Saffari, S.E. Application of the truncated zero-inflated double Poisson for determining of the effecting factors on the number of coronary artery stenosis. *Comput. Math. Methods Med.* **2022**, *1*, 5353539. [CrossRef] [PubMed]
46. Stasinopoulos, D.M.; Rigby, R.A. Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* **2008**, *23*, 1–46.
47. Stasinopoulos, M.D.; Kneib, T.; Klein, N.; Mayr, A.; Heller, G.Z. *Generalized Additive Models for Location, Scale and Shape: A Distributional Regression Approach, with Applications*; Cambridge University Press: Cambridge, UK, 2024; Volume 56.
48. Carroll, R.; Lawson, A.B.; Faes, C.; Kirby, R.S.; Aregay, M.; Watjou, K. Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. *Spat. Spatiotemporal Epidemiol.* **2015**, *14*, 45–54.
49. Akib, M.M.H.; Afroz, F.; Pal, B. Beyond averages: Dissecting urban-rural disparities in skilled antenatal care utilization in Bangladesh-a conway-maxwell-poisson regression analysis. *BMC Pregnancy Childbirth* **2025**, *25*, 119. [CrossRef]
50. Hohberg, M.; Pütz, P.; Kneib, T. Treatment effects beyond the mean using distributional regression: Methods and guidance. *PLoS ONE* **2020**, *15*, e0226514. [CrossRef]
51. Nogueira-Pileggi, V.; Achcar, M.C.; Carmona, F.; da Silva, A.C.; Aragon, D.C.; da Veiga Ued, F.; de Oliveira, M.M.; Fonseca, L.M.M.; Alves, L.G.; Bomfim, V.S.; et al. LioNeo project: A randomised double-blind clinical trial for nutrition of very-low-birth-weight infants. *Br. J. Nutr.* **2022**, *128*, 2490–2497.
52. Sathyanarayana, S.; Mohanasundaram, T. Fit indices in structural equation modeling and confirmatory factor analysis: Reporting guidelines. *Asian J. Econ. Bus. Account.* **2024**, *24*, 561–577. [CrossRef]
53. Sellers, K.F.; Morris, D.S. Underdispersion models: Models that are "under the radar". *Commun. Stat. Theory Methods* **2017**, *46*, 12075–12086. [CrossRef]
54. Lindsey, J.K.; Altham, P.M.E. Analysis of the human sex ratio by using overdispersion models. *J. R. Stat. Soc. C Appl. Stat.* **1998**, *47*, 149–157. [CrossRef] [PubMed]
55. Chang, H.Y.; Suchindran, C.M.; Pan, W.H. Using the overdispersed exponential family to estimate the distribution of usual daily intakes of people aged between 18 and 28 in Taiwan. *Stat. Med.* **2001**, *20*, 2337–2350. [CrossRef] [PubMed]
56. King, G. Variance specification in event count models: From restrictive assumptions to a generalized estimator. *Am. J. Political Sci.* **1989**, *33*, 762–784. [CrossRef]
57. Nandi, A.; Hazarika, P.J.; Biswas, A.; Hamedani, G.G. A new three-parameter discrete distribution to model over-dispersed count data. *Pak. J. Stat. Oper. Res.* **2024**, *20*, 197–215. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.