Biometrical Journal (20xx) , zzz–zzz / DOI: 10.1002/bimj.200100000

Zero–inflated regression models for radiation–induced chromosome aberration data: A comparative study

María Oliveira [1], **Jochen Einbeck**[*,1], **Manuel Higueras** [2,3], **Elizabeth Ainsbury** [2], **Pedro Puig** [3], and **Kai Rothkamm** [2,4]

[1] Department of Mathematical Sciences, Durham University, Durham DH1 3LE, UK
[2] Public Health England Centre for Radiation, Chemical and Environmental Hazards (PHE CRCE), Chilton, Didcot, Oxon OX11 0RQ, UK
[3] Departament de Matematiques, Universitat Autonoma de Barcelona, Bellaterra (Barcelona) 08193, Spain
[4] University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany

Received zzz, revised zzz, accepted zzz

Within the field of cytogenetic biodosimetry, Poisson regression is the classical approach for modelling the number of chromosome aberrations as a function of radiation dose. However, it is common to find data that exhibit overdispersion. In practice, the assumption of equidispersion may be violated due to unobserved heterogeneity in the cell population, which will render the variance of observed aberration counts larger than their mean, and/or the frequency of zero counts greater than expected for the Poisson distribution. This phenomenon is observable for both full and partial body exposure, but more pronounced for the latter. In this work, different methodologies for analysing cytogenetic chromosomal aberrations datasets are compared, with special focus on zero–inflated Poisson and zero–inflated negative binomial models. A score test for testing for zero–inflation in Poisson regression models under the identity link is also developed.

Key words: Biological dosimetry; Chromosome aberrations; Count data; Overdispersion; Zero–inflation; Score tests

Additional supporting information may be found in the online version of this article at the publisher's web-site.

1 Introduction

Data from biological systems regarding the effects of environmental or manmade mutagens frequently consist of count variables. This is the case in biological dosimetry, where the measurement of chromosome aberration frequencies in human lymphocytes is used for assessing absorbed doses of ionising radiation to individuals. For that purpose, dose–effect calibration curves are required which are produced by irradiating peripheral blood lymphocytes to a range of doses and quantifying the amount of damage induced by radiation at a cellular level, for instance by counting dicentrics or micronuclei (IAEA, 2011). That is, d blood samples from a healthy donor are irradiated with several doses x_i, $i = 1 \ldots, d$. Then for each irradiated sample, n_i cells are examined and the number of observed chromosomal aberrations y_{ij}, $j = 1, \ldots, n_i$ is recorded. The aberrations most commonly analyzed are the dicentrics, centric rings, and micronuclei.

These chromosomal aberrations appear because when cells are exposed to radiation, breaks are induced in the chromosomal DNA, and the broken fragments may rejoin incorrectly. Therefore, the frequency of chromosome aberrations increases with the amount of radiation and is a reliable and very well established biological indicator of radiation absorbed dose. Dicentrics are the interchange between the fragments of two separate chromosomes resulting in unstable, aberrant chromosomes with two centromeres. A ring chromosome, or centric ring, is an exchange between two breaks on separate arms of the same chromosome

*Corresponding author: e-mail: jochen.einbeck@durham.ac.uk, Phone: +44-191-3343125, Fax: +44-191-3343051

and is also accompanied by an acentric fragment (chromosome without centromere). Micronuclei are lagging chromosomal fragments or whole chromosomes at anaphase that are not included in the nuclei of daughter cells.

For such count data, the Poisson distribution is the most widely recognized and commonly used distribution and constitutes the standard framework for explaining the relationship between the outcome variable and the dose (Lloyd and Edwards, 1983; IAEA, 2011). However, in practice, the assumption of equidispersion implicit in the Poisson distribution is often violated, which is a well–known effect under high LET (Linear Energy Transfer) radiation, also known as densely ionising radiation (IAEA, 2011). Moreover, the distributions of micronuclei are in general overdispersed for both high and low LET radiation exposure.

The focus of the research presented in this manuscript is the identification of adequate response distributions for the modelling of cytogenetic dose–response curves. The cytogenetic dose estimation is a subsequent inverse regression problem that depends on this previous curve fitting. If the initial response distribution is incorrectly specified, this will impact on the accuracy of the model parameter estimates of the fitted curve and, more strongly, of their standard errors. In addition, the inverse regression step is sensitive to the initial model specification, and may behave unreliably if that specification is incorrect. Summarizing, an incorrectly specified response distribution may or may not lead to reasonable dose estimates, but it will certainly lead to an incorrect assessment of the uncertainty associated to these dose estimates. This subsequent inverse regression step is not the subject of this manuscript, see Higueras et al. (2015a, 2015b) for recent advances in this respect.

Due to the mentioned violations of the Poisson distribution, other distributions have been considered in the literature for dealing with overdispersed data in biodosimetry. These alternatives include the negative binomial distribution, which has been shown to accurately characterize aberration data in cases of overdispersion (Brame and Groer, 2002); the Neyman type A distribution, which has been shown to be useful for characterization of aberration induced by high LET radiation (Gudowska–Nowak *et al.*, 2007) and the univariate rth-order Hermite distributions (Puig and Barquinero, 2011). These distributions have recently been tested for suitability to a selection of chromosome aberration data collected in different exposure scenarios (Ainsbury *et al.*, 2013) and used for cytogenetic dose estimation through a Bayesian-like inverse regression technique (Higueras *et al.*, 2015a). Further, Poisson–inverse Gaussian and Pólya–Aeppli distributions have been considered in Puig and Valero (2006).

Also, a commonly observed characteristic of count data is the number of zeros in the sample exceeding the expected number of zeros generated by a Poisson distribution having the same mean. This phenomenon, known as zero–inflation, is frequently related to overdispersion. Distributions which account for overdispersion will also – to some extent – allow for zero-inflation. For instance, the families of Compound Poisson and Mixed Poisson distributions (which include the distributions mentioned in the previous paragraph as special cases) are overdispersed and zero-inflated.

However, the extra zeros (relative to the Poisson model) generated by these models may still be insufficient to account for the total observed number of zeros in the data. Count datasets with an excessive number of zero outcomes are abundant in many disciplines such as manufacturing applications (Lambert, 1992), medicine (Böhning *et al.*, 1999), econometrics (Gurmu *et al.*, 1999) and agriculture (Hall, 2000). In most of these works, a special kind of zero–inflated models are considered, using a mixture of a distribution degenerate at zero and a count distribution such a Poisson or a negative binomial. These models can be especially useful in partial body irradiation scenarios which feature a mixture of populations of non-irradiated and irradiated cells.

In this manuscript we will introduce and advocate the use of zero–inflated models for cytogenetic count data. We will compare zero–inflated models to other models previously proposed in the field of radiation biodosimetry, and we will devote particular attention to the question of whether overdispersion needs to be taken into account on top of the zero–inflation. The manuscript is organized as follows: In Section 2, zero–inflated Poisson and zero–inflated negative binomial models are reviewed. A case study involving several data sets with different radiation exposure patterns is provided in Section 3. In Section 4 we present a small simulation study in a radiation induced chromosome aberration context to study the identifiability of

zero–inflated and overdispersed regression models. The paper is concluded in Section 5. Appendix A.1 contains the derivations for a score test for zero–inflation under the identity link.

The code and data sets needed to reproduce the analyses carried out in this paper are available as supporting information. In addition, supplementary material is provided which gives the mathematical forms of count distributions used in Section 3, as well as further numerical results.

2 Zero–inflated regression models applied to biodosimetry

In this section, zero–inflated regression models are reviewed in a general framework in Section 2.1 and details on how these models are applied for modelling the number of chromosome aberrations as a function of radiation doses are given in Section 2.2.

2.1 Zero–inflated count regression overview

Zero–inflated count models provide one method to account for the excess zeros in data by modelling the data as a mixture of two distributions: a distribution taking a single value at zero and a count distribution such as Poisson or negative binomial distributions.

The zero–inflated Poisson (ZIP) regression model was first introduced by Lambert (1992) who applied the model to the data collected from a quality control study. Since then, the ZIP regression model has been applied in many and different fields, such as, dental epidemiology (Böhning *et al.*, 1999), occupational health (Lee *et al.*, 2001), and children's growth and development (Cheung 2002).

Let Y_{ij}, $i = 1, \ldots, d$, $j = 1, \ldots n_i$ be the response variable which in our context represents numbers of chromosomal aberrations at dose level i for cell j. A ZIP regression model is defined as

$$\mathrm{P}(Y_{ij} = y_{ij}) = \begin{cases} p_i + (1-p_i)\exp(-\lambda_i), & y_i = 0, \\ (1-p_i)\exp(-\lambda_i)\lambda_i^{y_{ij}}/y_{ij}!, & y_{ij} > 0, \end{cases}$$

where $0 \leq p_i \leq 1$ and $\lambda_i > 0$. For the ZIP, $\mathrm{E}(Y_{ij}) = (1-p_i)\lambda_i = \mu_i$ and $\mathrm{Var}(Y_{ij}) = (1-p_i)\lambda_i(1+p_i\lambda_i)$. Both the mean λ_i of the underlying Poisson distribution and the mixture parameter p_i (also referred to as 'zero–inflation parameter') can depend on vectors of covariates.

Since $\mathrm{Var}(Y_{ij}) = \mu_i(1 + p_i\lambda_i) \geq \mu_i$ it is clear that zero-inflation can be considered as a special form of overdispersion. When overdispersion is attributed to the large number of zeros with respect to the Poisson model, a ZIP model may provide a good fit. A ZIP model assumes that the zero observations have two different origins: some of them are zeros produced at random by the Poisson distribution, while some others (with proportion p_i) are "structural". The structural zeros have to be justified by the nature of data (in our case, by non–irradiated lymphocytes; for instance after partial body exposure). In addition, there may exist another source of overdispersion that cannot be attributed to the excess zeros. That is, even after accounting for zero–inflation, the non–zero part of the count distribution may be overdispersed (in our context, this will be mainly observed for densely ionising radiation). For dealing with this situation, Greene (1994) introduced an extended version of the negative binomial model for excess zero count data, the zero–inflated negative binomial (ZINB). In that case, when the overdispersion is both due to the heterogeneity of data and the excess of zeros, the ZINB regression model often is more appropriate than the ZIP.

For the ZINB regression model, the probability mass function of the response variable Y_{ij} ($i = 1, \ldots, d$, $j = 1, \ldots, n_i$) is given by

$$\mathrm{P}(Y_{ij} = y_{ij}) = \begin{cases} p_i + (1-p_i)(1+\alpha\lambda_i^c)^{-\lambda_i^{1-c}/\alpha}, & y_{ij} = 0, \\ (1-p_i)\frac{\Gamma(y_{ij}+\lambda_i^{1-c}/\alpha)}{y_{ij}!\Gamma(\lambda_i^{1-c}/\alpha)}(1+\alpha\lambda_i^c)^{-\lambda_i^{1-c}/\alpha}(1+\lambda_i^{-c}/\alpha)^{-y_{ij}}, & y_{ij} > 0, \end{cases}$$

where $\alpha > 0$ is an overdispersion parameter, and the index $c \in \{0, 1\}$ identifies the form of the underlying negative binomial distribution. These models will be denoted by ZINB1 and ZINB2, respectively. The

mean and variance of the ZINB distribution are $\mathrm{E}(Y_{ij}) = (1-p_i)\lambda_i = \mu_i$ and $\mathrm{Var}(Y_{ij}) = (1-p_i)\lambda_i(1 + p_i\lambda_i + \alpha\lambda_i^c)$, respectively. The ZINB model reduces to the ZIP model as $\alpha \to 0$, in analogy to the relationship between the negative binomial and the Poisson distribution.

2.2 Application to biological dosimetry

Count regression models such as Poisson and negative binomial and their zero–inflated versions have been widely applied in many and different fields. However their application to biological dosimetry deserves special attention.

In biodosimetry, it is assumed that the mean of the number of aberrations is a linear or a quadratic function of the dose (IAEA, 2011). For sparsely ionising radiation there is very strong evidence that the mean yield of chromosome aberrations, μ_i, is related to dose x_i by the quadratic equation:

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,\ i = 1, \ldots, d, \tag{1}$$

whereas for densely ionising radiation, the larger relative amount of energy deposited (and the increase in the density of the ionisations which lead to the damage measured) results in an increase in the linear term and the quadratic term becomes biologically less relevant and so, the dose response may be approximated by a linear equation.

The linear quadratic model is used for low linear-energy-transfer (LET) radiations (i.e. gamma and X-rays) based on the justification that dicentric chromosome aberrations and micronuclei result from interactions between two independently damaged chromosomes (Hall and Giaccia, 2012) and that the number of 'tracks' along which damage take place is linearly proportional to dose, so that the number of track (and thus damage) pairs is approximately proportional to dose squared (Hlatky *et al.*, 2002). For higher LET radiations, induction of chromosome aberrations becomes a linear function of dose because the more densely ionising nature of the radiation leads to a corresponding 'one track' distribution of damage. The same is true of fractionated or protracted doses, where there is time for repair of damage along one or more tracks between exposures.

Consequently, the link function used in (1) is simply the identity link function, as opposed to the log–link which is used for count data modelling in many other fields. The identity link is the accepted standard in biodosimetry since there is no evidence that the increase of aberration counts with dose is of exponential shape, and it avoids the undesired effect that dose–response curves start decreasing from about the maximum dose considered (IAEA, 2011). While we do not have strong arguments to change this standard, we point out that the log–link does have a few conceptual advantages, such as easier access to inferential tools for model testing, and the avoidance of problems with negative values of the linear predictor. In addition, using the log–link, that is,

$$\log(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,\ i = 1, \ldots, d, \tag{2}$$

a simple second order approximation of μ_i can be directly obtained applying Taylor's formula at $x_i = 0$,

$$\mu_i = \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \approx a + b x_i + c x_i^2\ i = 1, \ldots, d,$$

with $a = \exp(\beta_0)$, $b = \exp(\beta_0)\beta_1$ and $c = \exp(\beta_0)(\beta_2 + \beta_1^2/2)$. Therefore, for low doses the results obtained using the identity–link or the log–link have to be very similar. Indeed, we will find in our detailed study in the next section that the results obtained for the two link functions are largely interchangeable.

A consequence of using the identity–link is that the maximum likelihood estimate of the parameter β_0 obtained by maximizing the log–likelihood function of the corresponding model may be negative, i.e., may lead to a fitted negative control level, which makes no sense biologically. Therefore, in order to avoid negative values for the intercept, constraints in the domain of the parameters must be included when the model is fitted. Note that, though in some papers (e.g, Puig and Barquinero, 2011) the intercept is ignored in specific situations, it is well known that even when blood samples are not irradiated, the background

level of aberrations could be positive (IAEA, 2011). In the absence of an intercept, the likelihood function at dose 0 (and, hence, the full data likelihood) would take the value zero, for which reason we would advocate the general use of an intercept in any model. Furthermore, since radiation protection practises are generally very good these days, most 'real life' cytogenetic dose estimates are likely to be in the region of zero.

A decision is required on which mean function is to be modelled: the mean of the zero-inflated distribution, μ_i, or the mean of the underlying Poisson or negative binomial distribution, λ_i, which are related via $\lambda_i = \mu_i/(1-p_i)$. For compliance with formulation (1) and with practice in this particular field, we decided that it is adequate to model the mean of the corresponding zero–inflated distribution, μ_i, via the linear predictor in (1). If no covariates are assumed for p_i, then this is equivalent to modelling λ_i through a quadratic form.

The mixture parameter p_i will be modelled as usual with logistic regression, where three different scenarios will be investigated: Firstly, it is assumed that the proportion of the mixture is constant:

$$\text{logit}(p_i) = \gamma_0,\ i = 1, \ldots, d, \tag{3}$$

secondly, p_i is also modelled as a linear function of the dose:

$$\text{logit}(p_i) = \gamma_1 x_i,\ i = 1, \ldots, d, \tag{4}$$

and finally, p_i is also modelled as a linear function of the dose but an intercept is included:

$$\text{logit}(p_i) = \gamma_0 + \gamma_1 x_i,\ i = 1, \ldots, d. \tag{5}$$

These different approaches will be applied on several data sets in Section 3.3, and further discussed in Section 3.5.

It should be noted that the zero–inflated Poisson distribution has been previously applied to estimate the mean yield of aberrations of the irradiated fraction of cells and the dose received by this fraction by a patient who has been exposed to an inhomogeneous irradiation. This methodology, proposed by Dolphin (1969), is known in biodosimetry as Dolphin's method or contaminated Poisson method (IAEA, 2011). However, this methodology does not constitute 'zero–inflated regression' from the viewpoint of modern statistical modelling, as outlined in this section. So, while the concept of zero–inflation is not completely new in this context, at the best of our knowledge, zero–inflated *regression* models have not been employed for the construction of dose–response curves, neither for partial nor whole body exposure scenarios.

3 Comparative study

In order to study the performance of zero–inflated models to describe the number of chromosome aberrations in biological dosimetry a case study has been carried out where these models are compared with models already considered in the literature: Poisson, negative binomial, Neyman type A, Pólya–Aeppli and Poisson–inverse Gaussian. The mathematical forms of these distributions are given as supplementary material.

These models have been fitted following the "standards" given in Section 2.2 by using self–programmed code, which has been developed in the free software environment R (R Development Core Team, 2014). Function `maxLik` from package `maxLik` has been used in order to maximize the corresponding log–likelihood function. With the goal of facilitating the use of these techniques by practitioners, the function used for fitting the different models is available as supporting information jointly with a detailed description of its usage and the datasets used in the study.

3.1 Scenarios: description of the data

The models have been applied to several real datasets obtained under four different scenarios: whole and partial body exposure with sparsely and densely ionising radiation. A brief description of them is given below.

(A) Whole body exposure – sparsely ionising radiation:

(A1) These data consist of the frequency of dicentric chromosomes after acute whole body *in vitro* exposure to eight uniform doses between 0 and 4.5 Gy of Cobalt–60 gamma rays (dose rate: 0.27 Gy/min). Blood was taken from fourteen healthy donors (six for the 0 Gy controls, and eight for the irradiated samples). Data were collected within the MULTIBIODOSE project and can be found in Table 6 of Romm *et al.* (2013).

(A2) This dataset consists of scores of micronuclei obtained after irradiating eleven samples of peripheral blood with different doses (between 0 and 4 Gy) of gamma irradiation, where the dose rate was 0.93 cGy/min. In this case, for each sample, approximately 5000 binucleated cells were inspected and the numbers of micronuclei were counted. Data can be found in Table 6 of Puig and Valero (2006).

(A3) Frequencies of dicentrics and centric rings aberrations are analysed in a total of 51600 metaphases from two volunteers after whole body exposure with 200 kV X-rays. Data considered here were obtained by scoring in metaphases reaching the first mitosis after a culture time of 56 h. Data can be found in Table 2 of Heimers *et al.* (2006).

(B) Whole body exposure – densely ionising radiation:

(B1) This dataset corresponds to the number of dicentrics after exposure of peripheral blood samples to 10 different doses (between 0 and 1.6 Gy) of 1480 MeV oxygen ions. Data can be found in Table 2 of Di Giorgio *et al.* (2004) and was studied by Puig and Barquinero (2011).

(B2) The second dataset considered in this scenario was obtained after irradiating blood samples with five different doses between 0.1 and 1 Gy of 2.1 MeV neutrons. In this case, frequencies of dicentrics and centric rings are analysed. Data are from Table 3 from Heimers *et al.* (2006) and correspond to a culture time of 72 h.

(C) Partial body exposure – sparsely ionising radiation:

(C1–C3) Three datasets were considered here. The scenario is the same as for dataset (A3) but, they correspond to partial body exposure simulation, with unirradiated blood mixed with irradiated blood from the same donors. The proportion of irradiated blood is 25%, 50% and 75%, respectively.

(D) Partial body exposure – densely ionising radiation:

(D1–D3) Finally, three datasets are considered in this scenario. Data were obtained by irradiating blood samples with 2.1 MeV neutrons (as in B2) and the same culture time is considered. The proportion of irradiated blood is 25%, 50% and 75%, respectively. The data for both scenarios (C) and (D) are available from Heimers *et al.* (2006).

Quadratic dose models of type (1) and (2) will be used under sparsely ionising radiation, that is for data sets (A1) to (A3) and (C1) to (C3), and, following Puig and Barquinero (2011), also for data set (B1). Following the reasoning outlined in Section 2.2, the quadratic term term will be removed for data sets (B2) and (D1) to (D3).

To illustrate the nature of the data, the full data set (B1) is displayed in Table 1 and visualized in Figure 1. (Analogous tables and graphs for the remaining datasets are available as supporting information.) Recall

Table 1 Doses, frequency distributions of the number of dicentrics, sample size and sum, and u-test values, for data set (B1).

	y_{ij}										
x_i	0	1	2	3	4	5	6	7	n_i	y_i	u_i
0.000	1999	1	0	0	0	0	0	0	2000	1	0
0.092	737	16	0	0	0	0	0	0	753	16	-0.399
0.120	1438	55	5	2	0	0	0	0	1500	71	7.261
0.205	1300	104	14	2	0	0	0	0	1420	138	5.173
0.300	471	73	15	1	0	0	0	0	560	106	2.560
0.405	437	66	15	1	1	0	0	0	520	103	4.377
0.600	473	119	34	3	2	0	0	0	631	204	3.876
0.820	253	99	38	17	5	0	0	1	413	253	7.158
1.200	92	55	27	11	4	1	0	0	190	163	2.948
1.600	80	49	26	13	5	0	0	0	173	160	2.512

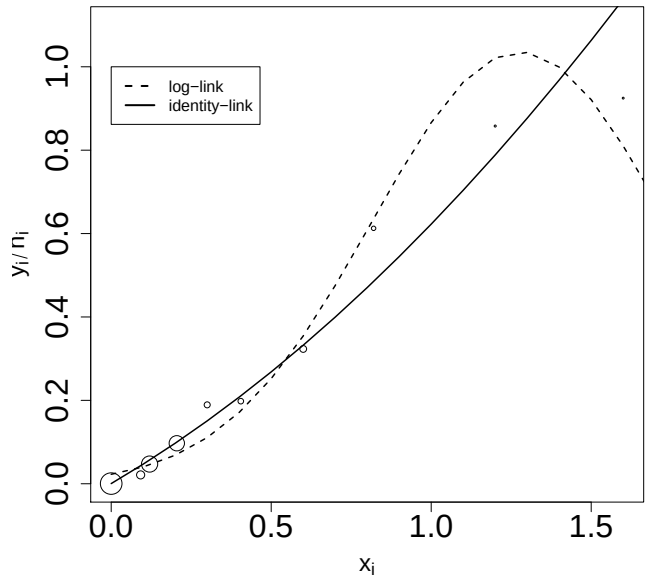

Figure 1 Dataset (B1): Proportions y_i/n_i (symbolized by circles of radius $\propto n_i$) and dose-response curves fitted with Poisson model and two link functions.

that we denote $y_i = \sum_{j=1}^{n_i} y_{ij}$ the total number of counts observed for dose x_i, that is, y_i is the sufficient statistics to estimate the mean of the Poisson distribution under dose x_i. In the graphical representation, the circles have location $(x_i, y_i/n_i)$ and size n_i. The solid curve is the dose–response curve that would be fitted according to the Poisson model with identity link. However, consider the u_i figures shown in Table 1 which are the values of the u-test statistic of Rao and Chakravarti (1956) to measure the overdispersion, suggested by IAEA (2011). Most of these u values are > 1.96 (except for the control and the 0.092 Gy samples), rejecting in general the equidispersion assumption, thus the classical Poisson model is not appropriate for fitting this dataset.

3.2 Score tests

Before we provide our detailed overview of fitted models, we will give some more evidence for the presence of zero–inflation, and overdispersion on top of the zero–inflation, in the datasets introduced in Section 3.1. Score (or Rao) tests are a convenient tool for this purpose. A score test for testing a Poisson against a ZIP regression model was developed by van den Broek (1995). Similar score tests do exist for testing a Poisson against a negative binomial (NB) regression model (Dean and Lawless, 1989), as well as as ZIP against a ZINB regression model (Ridout *et al*, 2001). All these tests assume that constant probabilities (3) are employed. Furthermore, all these tests require that the mean is modelled through a log–link function. For the Poisson/ZIP case, we developed a variant of van den Broek's score test which also works under the identity link; see the appendix for details. As one can see from Table 2, the values of the test statistic are quite similar for the two link functions, and in any case lead to the same conclusions.

Similar adaptions of the score test for the identity link could be developed for the Poisson/NB and the ZIP/ZINB comparisons though this is beyond the scope of this paper. Hence, for these two latter situations, we constrain ourselves to log–link models when applying the score–test (our considerations in Section 2.2, as well as the results of the Poisson/ZIP test, suggest that this is not a serious restriction).

The values of the score test statistics for all considered datasets are given in Table 2. The values given in this table are compared with quantiles of the chi–squared distribution with one degree of freedom; for instance at the 5% levels of significance this quantile takes the value 3.84 (see the last paragraph of the appendix for further discussion). The higher the provided value of the test statistics, the stronger the evidence against the smaller model. This leads to the following conclusions:

- only for dataset (A3) — sparsely ionising whole body exposure — the assumption of a Poisson distribution cannot be rejected;
- for dataset (A1) — again, sparsely ionising whole body exposure — the Poisson assumption is rejected;
- for all datasets involving densely ionising radiation, that is (B) and (D), as well as for the micronuclei (A2), the Poisson model is rejected in favour of the ZIP and NB models, and furthermore the ZIP model is rejected in favour of the ZINB model.
- for all data sets involving partial body exposure, that is (C) — sparsely ionising — and (D) — densely ionising —, the Poisson assumption is rejected in favour of the ZIP and NB models.

It is worth noting that the current IAEA recommendation is for the Poisson distribution to be applied to all sparsely ionising data, testing for overdispersion then applying Dolphin's (1969) contaminated Poisson method or similar when Poisson assumptions are violated — which is expected in partial body exposure scenarios (IAEA, 2011). This recommendation is not entirely at odds with the result of our initial score tests, but is clearly too vague to be actually useful, so that cytogenists, in the absence of further guidance, tend to use the — apparently incorrect — Poisson assumption in most of the cases.

From the score test results one can further observe that, while for some datasets it will be sufficient to model either overdispersion or zero–inflation, for other datasets overdispersion appears to be separately present on top of the zero–inflation. We continue with a comprehensive analysis, fitting these and a variety of other related models, which will confirm these results.

3.3 Results

In order to compare the performance of the different models, classical likelihood measures of goodness of fit are used: The Akaike Information Criterion (AIC) and the Bayesian (Schwarz) Information Criterion (BIC). The AIC (Akaike, 1974) penalizes a model with a larger number of parameters, and is defined as $\mathrm{AIC} = -2\log L + 2k$, where $\log L$ denotes the fitted log–likelihood and k the number of parameters.

Table 2 Values of the score test statistic for data sets (A1)–(D3) and several test problems, with 'P' denoting 'Poisson'. The form of the (zero–inflated) negative binomial model considered in each case is the one that provided the best fit according to the log–likelihood value in Tables 3 to 6. For tests involving zero–inflated models, the mixture parameter has been modelled according to (3).

link	test	(A1)	(A2)	(A3)	(B1)	(B2)	(C1)	(C2)	(C3)	(D1)	(D2)	(D3)
id	P/ZIP	18.17	383.58	0.92	87.72	61.32	2007.39	1418.28	776.55	416.20	387.91	168.13
	P/ZIP	16.89	378.69	1.00	87.16	47.20	1996.30	1417.96	745.84	421.48	398.38	168.74
log	P/NB	20.79	1699.91	0.90	159.26	136.89	6009.35	3281.00	1210.34	770.62	693.80	285.61
	ZIP/ZINB	1.54	1043.94		47.20	64.96	0.22	1.74	0.01	11.49	35.94	36.24

The BIC (Schwarz, 1978), defined as $\mathrm{BIC} = -2\log L + k\log n$, works similarly to AIC but increases the penalty with increasing sample size n (with our notation $n = \sum_{i=1}^{d} n_i$). According to these criteria, models with smaller values of AIC and BIC are considered preferable. It is standard practise to include both criteria in model fitting. Tables 3–6 show the results for each dataset considered, for both the identity link and the log–link (first and second row, respectively, for each given model). The value k used for AIC and BIC is given explicitly in each table, and is computed as the sum of regression and model parameters. The best model in each column and for each link function is provided in bold face.

(A) Whole body exposure – sparsely ionising radiation

Firstly, we observe from Table 3 that, as expected from the result of the score test, for dataset (A3) the Poisson model comes up as the preferred model under both the AIC and the BIC criterion. This corresponds to accepted practice for dicentrics under whole body exposure and sparsely ionising radiation.

However, for dataset (A1), the values of the maximized log–likelihood as well as the information criteria indicate that NB2 and zero–inflated models fit the data better than other models. A similar behavior has been reported for other datasets in the literature (e.g., for data corresponding to lab 3 shown in Table 3 in Romm et al., 2013) obtained using an automatic scoring procedure. In this case, one could speculate that the automatic scoring procedure used for (A1) may skew the data away from Poisson. However, more datasets would be needed to demonstrate such an effect reliably.

For data (A2), the Poisson distribution does not provide a good fit (see Table 3). In this case, it should be pointed out that micronuclei counts differ from dicentrics in that i) the quadratic component of the dose dependence is frequently weaker (for sparsely irradiation), ii) baseline counts of unirradiated samples are much higher than for dicentrics and iii) even after uniform total body irradiation micronucleus distributions tend to be overdispersed.

Therefore, although for whole–body exposure and sparsely ionising radiation, it is usually assumed that data follow a Poisson model, data under this scenario may depart from the Poisson model due to other circumstances (e.g., the scoring procedure).

(B) Whole body exposure – densely ionising radiation

For the two datasets in this scenario values for the Poisson regression model are clearly worse than for the other models, confirming the overdispersion reported for several authors concerning high LET radiation exposures. According to the results shown in Table 4, there are several models which are very competitive. In this case, it seems that overdispersion can be modelled through different models, including the Neyman A and zero–inflated negative binomial models.

(C) Partial body exposure – sparsely ionising radiation

For datasets considered in this scenario (C1–C3), zero–inflated models are notably better than the other models as shown in Table 5. This result is in line with the philosophy of Dolphin's method (Dolphin, 1969). The zero–inflated Poisson models perform consistently well for all three datasets, and the information criteria give little support for (possibly zero–inflated) negative binomial models. Hence, for this type of datasets, it seems clear that overdispersion is due to the excess of zeros.

(D) Partial body exposure – densely ionising radiation

Table 3 Results of fitting various models to datasets (A1), (A2) and (A3), obtained under whole body exposure with sparsely ionising radiation. For each model, results obtained with identity–link (first row) and log–link (second row, *italic*) are shown.

		(A1)			(A2)			(A3)		
Models	k	loglik	AIC	BIC	loglik	AIC	BIC	loglik	AIC	BIC
Poisson	3	-3748.59	7503.17	7526.14	-34679.48	69364.95	69391.70	-3806.89	**7619.77**	**7638.94**
	3	*-3749.36*	*7504.73*	*7527.70*	*-34721.03*	*69448.07*	*69474.81*	*-3808.27*	***7622.55***	***7641.72***
NB1	4	-3742.82	7493.65	7524.28	-34199.14	68406.28	68441.94	-3806.89	7621.77	7647.33
	4	*-3743.69*	*7495.39*	*7526.02*	*-34231.24*	*68470.47*	*68506.13*	*-3808.28*	*7624.55*	*7650.11*
NB2	4	-3739.23	**7486.46**	**7517.09**	-34398.42	68804.84	68840.50	-3806.98	7621.97	7647.53
	4	*-3740.55*	***7489.10***	***7519.73***	*-34440.88*	*68889.76*	*68925.42*	*-3808.28*	*7624.56*	*7650.12*
Neyman A	4	-3742.96	7493.91	7524.54	-34214.13	68436.25	68471.91	-3806.93	7621.86	7647.42
	4	*-3743.78*	*7495.57*	*7526.20*	*-34246.64*	*68501.27*	*68536.93*	*-3808.30*	*7624.60*	*7650.16*
Polya-Aeppli	4	-3742.90	7493.79	7524.42	-34204.54	68417.09	68452.75	-3806.93	7621.86	7647.42
	4	*-3743.74*	*7495.47*	*7526.10*	*-34236.76*	*68481.51*	*68517.17*	*-3808.31*	*7624.63*	*7650.18*
PIG	4	-3742.75	7493.50	7524.13	-34196.84	68401.69	**68437.35**	-3806.91	7621.82	7647.38
	4	*-3743.62*	*7495.25*	*7525.88*	*-34228.97*	*68465.94*	***68501.60***	*-3808.28*	*7624.56*	*7650.12*
ZIP (3)	4	-3739.79	7487.58	7518.21	-34490.47	68988.94	69024.60	-3806.44	7620.87	7646.43
	4	*-3741.18*	*7490.36*	*7520.99*	*-34534.06*	*69076.12*	*69111.78*	*-3807.78*	*7623.57*	*7649.12*
ZIP (4)	4	-3741.26	7490.52	7521.15	-34352.76	68713.53	68749.19	-3806.89	7621.78	7647.34
	4	*-3742.85*	*7493.69*	*7524.32*	*-34395.01*	*68798.02*	*68833.68*	*-3808.28*	*7624.55*	*7650.11*
ZIP (5)	5	-3739.18	7488.36	7526.65	-34266.33	68542.66	68587.23	**-3806.21**	7622.41	7654.36
	5	*-3740.19*	*7490.38*	*7528.67*	*-34299.43*	*68608.87*	*68653.44*	***-3807.55***	*7625.11*	*7657.06*
ZINB1 (3)	5	-3739.69	7489.38	7527.67	-34199.16	68408.31	68452.89	-3806.45	7622.91	7654.85
	5	*-3740.72*	*7491.44*	*7529.73*	*-34231.63*	*68473.26*	*68517.84*	*-3808.19*	*7626.39*	*7658.33*
ZINB1 (4)	5	-3741.27	7492.53	7530.82	**-34195.50**	**68400.99**	68445.57	-3807.24	7624.49	7656.43
	5	*-3742.81*	*7495.62*	*7533.90*	*-34226.81*	*68463.62*	*68508.19*	*-3808.38*	*7626.76*	*7658.71*
ZINB1 (5)	6	-3742.82	7497.65	7543.59	-34195.73	68403.46	68456.96	-3807.03	7626.06	7664.40
	6	*-3739.38*	*7490.75*	*7536.70*	***-34224.79***	***68461.58***	*68515.07*	*-3808.31*	*7628.62*	*7666.95*
ZINB2 (3)	5	-3739.14	7488.27	7526.56	-34398.60	68807.20	68851.78	-3806.44	7622.87	7654.82
	5	*-3740.49*	*7490.98*	*7529.26*	*-34440.92*	*68891.84*	*68936.41*	*-3807.78*	*7625.57*	*7657.51*
ZINB2 (4)	5	-3739.08	7488.16	7526.45	-34281.79	68573.58	68618.16	-3806.89	7623.78	7655.73
	5	*-3740.37*	*7490.74*	*7529.03*	*-34322.27*	*68654.54*	*68699.12*	*-3815.15*	*7640.30*	*7672.25*
ZINB2 (5)	6	**-3738.15**	7488.30	7534.25	-34210.50	68433.00	68486.49	-3806.21	7642.42	7662.75
	6	***-3739.25***	*7490.50*	*7536.45*	*-34242.98*	*68497.96*	*68551.45*	*-3807.55*	*7627.11*	*7665.44*

For datasets in this scenario (D1–D3), the Poisson model is clearly rejected. From Table 6, it can be observed that, in general, the ZINB models provide the best fits which indicates that overdispersion is due to both the excess of zeroes (caused by the partial body exposure) and the heterogeneity (caused by the densely ionising radiation). However, there is quite a wide range of models which provided competitive results for some data sets under this scenario, among them NB2, Polya-Aeppli, and the Neyman type A model. The latter has been shown to perform well for densely ionising radiation by Virsik and Harder (1981).

In our analysis, the Poisson model provides the (by far) worst fit for almost all datasets, including the sparsely ionising scenarios. Thus, a Poisson model should be used only in cases where there is strong evidence that it is the correct specification. In any case, it is clear that the Poisson model will be inadequate under partial body exposure and/or for densely ionising radiation. In general, as compared to the Poisson model, the proposed zero–inflated regression models perform well in terms of log–likelihood and the model selection criteria employed, for both full and partial body exposure.

Table 4 Results of fitting various models to datasets (B1) and (B2), obtained under whole body exposure with densely ionising radiation. For each model, results obtained with identity–link (first row) and log–link (second row, *italic*) are shown. Separate columns for k are provided for dataset (B1), which employs a quadratic model, and dataset (B2), which uses a linear predictor without quadratic term.

		(B1)			(B2)			
Models	k	loglik	AIC	BIC	loglik	AIC	BIC	k
Poisson	3	-2855.85	5717.70	5738.73	-3004.72	6013.45	6026.57	2
	3	*-2904.50*	*5815.00*	*5836.02*	*-3028.27*	*6060.54*	*6073.66*	*2*
NB1	4	-2800.29	5608.58	5636.60	-2960.18	5926.36	5946.05	3
	4	*-2846.15*	*5700.30*	*5728.33*	*-2977.92*	*5961.83*	*5981.52*	*3*
NB2	4	-2807.48	5622.96	5650.99	-2976.17	5958.34	5978.03	3
	4	*-2856.61*	*5721.22*	*5749.25*	*-2996.11*	*5998.22*	*6017.91*	*3*
Neyman A	4	-2799.74	5607.47	**5635.50**	-2958.86	**5923.72**	**5943.40**	3
	4	*-2845.21*	*5698.41*	*5726.44*	*-2976.94*	***5959.88***	***5979.57***	*3*
Polya-Aeppli	4	-2799.81	5607.61	5635.64	-2959.41	5924.81	5944.50	3
	4	*-2845.48*	*5698.97*	*5727.00*	*-2977.25*	*5960.50*	*5980.19*	*3*
PIG	4	-2801.91	5611.81	5639.84	-2961.99	5929.97	5949.66	3
	4	*-2848.04*	*5704.08*	*5732.11*	*-2979.74*	*5965.48*	*5985.17*	*3*
ZIP (3)	4	-2814.53	5637.07	5665.09	-2979.09	5964.19	5983.88	3
	4	*-2861.85*	*5731.69*	*5759.72*	*-3005.82*	*6017.64*	*6037.33*	*3*
ZIP (4)	4	-2805.36	5618.71	5646.74	-2967.53	5941.05	5960.74	3
	4	*-2854.06*	*5716.12*	*5744.15*	*-2990.43*	*5986.87*	*6006.56*	*3*
ZIP (5)	5	-2800.58	5611.17	5646.20	-2958.51	5925.01	5951.26	4
	5	*-2847.77*	*5705.53*	*5740.57*	*-2977.43*	*5962.86*	*5989.12*	*4*
ZINB1 (3)	5	-2797.41	5604.82	5639.85	-2961.15	5930.31	5956.56	4
	5	*-2842.31*	*5694.63*	*5729.66*	*-2977.92*	*5963.84*	*5990.09*	*4*
ZINB1 (4)	5	**-2797.30**	**5604.61**	5639.64	-2962.50	5933.00	5959.25	4
	5	*-2842.34*	*5694.68*	*5729.72*	*-2976.85*	*5961.70*	*5987.95*	*4*
ZINB1 (5)	6	-2797.33	5606.67	5648.71	**-2957.62**	5925.25	5958.06	5
	6	*-2842.04*	*5696.07*	*5738.11*	**-2975.95**	*5961.90*	*5994.71*	*5*
ZINB2 (3)	5	-2807.47	5624.93	5659.97	-2976.03	5960.06	5986.32	4
	5	*-2856.41*	*5722.82*	*5757.86*	*-2996.13*	*6000.26*	*6026.51*	*4*
ZINB2 (4)	5	-2800.06	5610.13	5645.16	-2964.15	5936.29	5962.54	4
	5	*-2847.84*	*5705.68*	*5740.71*	*-2984.50*	*5976.99*	*6003.24*	*4*
ZINB2 (5)	6	-2798.59	5609.17	5651.22	-2957.84	5925.68	5958.49	5
	6	***-2809.61***	***5631.21***	***5673.25***	*-2976.29*	*5962.58*	*5995.40*	*5*

3.4 Alternative model classes

The wide range of model classes considered so far does not make the claim to be exhaustive, and there are further modelling strategies which deserve consideration.

Firstly, from a conceptual viewpoint, Hermite regression models provide an attractive class of models in biodosimetry. If the radioactive particles arriving to the cell follow a Poisson process and each particle can produce simultaneously up to r dicentrics, then the resulting distribution is just a Hermite distribution of order r (Puig and Barquinero, 2011). Note that the order $r = 1$ corresponds just to the Poisson distribution, results for which can be read from Tables 3 to 6. We have provided results for Hermite models of order $r = 2, 3$ and 4 in Table 1 of the supplementary material. One finds that, for whole body exposure, Hermite models with $r = 2$ performed competitively to other models discussed earlier. For partial body exposure, where one would consider orders $r \geq 3$, Hermite models were more successful under scenario (D) (high LET) than under (C). We do not consider Hermite models with $r \geq 5$, as these would be hard to justify, and would possess an excessively large amount of parameters.

Table 5 Results of fitting various models to datasets (C1), (C2) and (C3), obtained under partial body exposure with sparsely ionising radiation. For each model, results obtained with identity–link (first row) and log–link (second row, *italic*) are shown.

		(C1)			(C2)			(C3)		
Models	k	loglik	AIC	BIC	loglik	AIC	BIC	loglik	AIC	BIC
Poisson	3	-2674.93	5355.86	5376.50	-3526.90	7059.81	7079.70	-3472.24	6950.47	6969.64
	3	*-2676.09*	*5358.18*	*5378.83*	*-3528.70*	*7063.39*	*7083.28*	*-3468.15*	*6942.30*	*6961.46*
NB1	4	-2090.11	4188.21	4215.74	-3011.85	6031.70	6058.23	-3229.20	6466.40	6491.95
	4	*-2091.83*	*4191.65*	*4219.18*	*-3011.69*	*6031.38*	*6057.90*	*-3224.49*	*6456.98*	*6482.54*
NB2	4	-2088.53	4185.07	4212.59	-2939.48	5886.97	5913.49	-3155.36	6318.71	6344.27
	4	*-2052.98*	*4113.96*	*4141.48*	*-2940.52*	*5889.05*	*5915.57*	*-3153.54*	*6315.08*	*6340.64*
Neyman A	4	-2103.10	4214.20	4241.73	-3021.07	6050.13	6076.66	-3232.00	6471.99	6497.55
	4	*-2104.75*	*4217.50*	*4245.03*	*-3022.38*	*6052.76*	*6079.28*	*-3229.16*	*6466.31*	*6491.87*
Polya-Aeppli	4	-2087.21	4182.42	4209.94	-3007.02	6022.04	6048.56	-3227.37	6462.75	6488.31
	4	*-2088.91*	*4185.82*	*4213.35*	*-3007.94*	*6023.89*	*6050.41*	*-3223.57*	*6455.14*	*6480.69*
PIG	4	-2109.59	4227.19	4254.72	-3035.89	6079.79	6106.31	-3241.24	6490.47	6516.03
	4	*-2111.35*	*4230.69*	*4258.22*	*-3035.98*	*6079.96*	*6106.48*	*-3235.40*	*6478.81*	*6504.37*
ZIP (3)	4	-2010.84	4029.68	***4057.21***	-2852.63	5713.26	5739.79	-3092.40	6192.79	6218.35
	4	*-2010.76*	*4029.53*	***4057.05***	*-2852.29*	*5712.59*	*5739.11*	*-3092.73*	*6193.46*	*6219.02*
ZIP (4)	4	-2034.75	4077.51	4105.03	-2844.89	5697.77	**5724.29**	-3087.70	6183.41	***6208.97***
	4	*-2026.52*	*4061.05*	*4088.58*	*-2845.72*	*5699.44*	***5725.96***	*-3086.79*	*6181.58*	*6207.14*
ZIP (5)	5	-2007.01	***4024.02***	4058.43	**-2842.39**	**5694.79**	5727.94	-3085.56	***6181.13***	6213.08
	5	*-2006.57*	***4023.13***	*4057.54*	***-2843.70***	***5697.40***	*5730.55*	*-3081.33*	***6172.66***	***6204.61***
ZINB1 (3)	5	-2010.85	4031.70	4066.10	-2852.65	5715.31	5748.46	-3092.45	6194.91	6226.85
	5	*-2010.78*	*4031.55*	*4065.96*	*-2852.31*	*5714.61*	*5747.77*	*-3092.75*	*6195.50*	*6227.44*
ZINB1 (4)	5	-2017.66	4045.31	4079.72	-2844.21	5698.43	5731.58	-3087.84	6185.67	6217.62
	5	*-2015.63*	*4041.25*	*4075.66*	*-2845.06*	*5700.13*	*5733.28*	*-3086.80*	*6183.59*	*6215.54*
ZINB1 (5)	6	-2006.97	4025.94	4067.23	-2842.41	5696.81	5736.60	-3085.55	6183.10	6221.44
	6	*-2006.50*	*4024.99*	*4066.28*	*-2843.71*	*5699.42*	*5739.20*	*-3080.99*	*6173.97*	*6212.31*
ZINB2 (3)	5	-2010.70	4031.40	4065.81	-2852.63	5715.26	5748.42	-3092.37	6194.74	6226.68
	5	*-2010.66*	*4031.31*	*4065.72*	*-2852.29*	*5714.59*	*5747.74*	*-3092.73*	*6195.46*	*6227.40*
ZINB2 (4)	5	-2022.00	4054.00	4088.41	-2844.88	5699.75	5732.90	-3087.68	6185.37	6217.32
	5	*-2021.05*	*4052.10*	*4086.50*	*-2845.72*	*5701.44*	*5734.59*	*-3086.79*	*6183.58*	*6215.53*
ZINB2 (5)	6	***-2006.47***	4024.93	4066.22	-2842.42	5696.83	5736.62	***-3084.93***	6181.87	6220.20
	6	***-2006.20***	*4024.41*	*4065.70*	*-2843.70*	*5699.40*	*5739.18*	***-3080.94***	*6173.87*	*6212.21*

Secondly, we investigated random effect models which add a random intercept term to the linear predictor (1) or (2). Considering the responses as repeated measures y_{ij} equipped with a two–level structure, the random effect is added to the upper (aggregated) data level, i, effectively imposing correlation within blood samples. While correlation between cells from the same blood sample is a reasonable assumption, we note that each blood sample got exposed to a different dose, which is included as covariate into the model. We certainly would expect the dose effect to be much larger than any possible within–sample correlation. Hence, we do not consider this approach as a truly hierarchical ('variance component') model, but rather as a simple overdispersion model. We fitted random effect models using a Gaussian random effect with Poisson and negative binomial response distribution, and, for the former, also considered an unspecified random effect distribution. Detailed results are provided in Table 2 of the supplementary material. Encouragingly, for data sets (A1) and (D2), the negative binomial random effect model using the log–link produced better results (in terms of BIC) than any of the previously discussed models. However, the random effect models did not perform uniformly well, and at some occasions run into computational difficulties. The identity–link version, for which we found a workable implementation only for the Poisson model with Gaussian random effect, is more difficult to use than the log–link since the random effect can render the linear predictor negative, which is incompatible with its interpretation as a Poisson mean. Of course, random effect models will show their actual power only in truly hierarchical setups, where they can

Table 6 Results of fitting various models to datasets (D1), (D2) and (D3), obtained under partial body exposure with densely ionising radiation. For each model, results obtained with identity–link (first row) and log–link (second row, *italic*) are shown.

		(D1)			(D2)			(D3)		
Models	k	loglik	AIC	BIC	loglik	AIC	BIC	loglik	AIC	BIC
Poisson	2	-1477.95	2959.89	2973.54	-2302.09	4608.18	4621.51	-2394.99	4793.98	4806.93
	2	*-1482.50*	*2969.00*	*2982.65*	*-2323.30*	*4650.60*	*4663.94*	*-2415.08*	*4834.17*	*4847.12*
NB1	3	-1370.07	2746.13	2766.61	-2148.66	4303.31	4323.31	-2310.45	4626.89	4646.32
	3	*-1373.15*	*2752.30*	*2772.77*	*-2163.63*	*4333.26*	*4353.26*	*-2326.58*	*4659.17*	*4678.59*
NB2	3	-1366.28	**2738.57**	**2759.04**	-2151.63	4309.26	4329.26	-2322.30	4650.59	4670.02
	3	*-1370.12*	***2746.25***	***2766.72***	*-2167.41*	*4340.81*	*4360.81*	*-2337.16*	*4680.31*	*4699.74*
Neyman A	3	-1372.33	2750.65	2771.13	-2146.95	4299.91	4319.91	-2306.20	**4618.39**	**4637.82**
	3	*-1375.41*	*2756.81*	*2777.29*	*-2161.79*	*4329.58*	*4349.58*	*-2322.23*	***4650.47***	***4669.89***
Polya-Aeppli	3	-1370.34	2746.67	2767.15	-2146.66	4299.32	**4319.32**	-2308.06	4622.11	4641.54
	3	*-1373.41*	*2752.83*	*2773.30*	*-2161.53*	*4329.06*	***4349.06***	*-2324.10*	*4654.20*	*4673.63*
PIG	3	-1371.72	2749.43	2769.90	-2155.04	4316.09	4336.08	-2315.55	4637.11	4656.54
	3	*-1374.81*	*2755.63*	*2776.10*	*-2170.28*	*4346.57*	*4366.57*	*-2331.96*	*4669.93*	*4689.36*
ZIP (3)	3	-1369.48	2744.96	2765.43	-2155.15	4316.30	4336.30	-2322.37	4650.73	4670.16
	3	*-1373.58*	*2753.17*	*2773.64*	*-2173.27*	*4352.53*	*4372.53*	*-2341.29*	*4688.58*	*4708.01*
ZIP (4)	3	-1386.57	2779.15	2799.62	-2172.81	4351.62	4371.61	-2323.33	4652.66	4672.09
	3	*-1391.84*	*2789.68*	*2810.16*	*-2193.54*	*4393.07*	*4413.07*	*-2341.86*	*4689.72*	*4709.15*
ZIP (5)	4	-1368.96	2745.91	2773.21	-2147.03	4302.07	4328.73	-2308.05	4624.11	4650.01
	4	*-1372.62*	*2753.25*	*2780.55*	*-2160.67*	*4329.34*	*4356.00*	*-2321.58*	*4651.15*	*4677.06*
ZINB1 (3)	4	-1366.48	2740.96	2768.26	**-2143.46**	**4294.93**	4321.59	-2308.53	4625.06	4650.97
	4	*-1369.76*	*2747.53*	*2774.83*	*-2158.76*	***4325.53***	*4352.19*	*-2324.87*	*4657.73*	*4683.64*
ZINB1 (4)	4	-1366.16	2740.32	2767.62	-2143.59	4295.19	4321.85	-2307.72	4623.44	4649.35
	4	*-1373.16*	*2754.32*	*2781.61*	*-2158.79*	*4325.59*	*4352.25*	*-2323.98*	*4655.97*	*4681.87*
ZINB1 (5)	5	-1366.13	2742.27	2776.39	-2143.40	4296.80	4330.13	-2306.96	4623.93	4656.31
	5	***-1369.22***	*2748.45*	*2782.57*	***-2158.67***	*4327.34*	*4360.66*	*-2321.48*	*4652.96*	*4685.35*
ZINB2 (3)	4	-1366.05	2740.09	2767.39	-2150.67	4309.35	4336.01	-2320.61	4649.22	4675.13
	4	*-1369.93*	*2747.85*	*2775.15*	*-2166.94*	*4341.88*	*4368.54*	*-2336.76*	*4681.52*	*4707.42*
ZINB2 (4)	4	-1366.44	2740.87	2768.17	-2147.31	4302.62	4329.28	-2313.89	4635.79	4661.69
	4	*-1369.97*	*2747.94*	*2775.23*	*-2162.26*	*4332.53*	*4359.19*	*-2328.61*	*4665.22*	*4691.12*
ZINB2 (5)	5	**-1365.88**	2741.77	2775.89	-2144.92	4299.85	4333.18	**-2307.69**	4625.38	4657.76
	5	*-1369.66*	*2749.32*	*2783.45*	*-2158.93*	*4327.85*	*4361.18*	***-2321.34***	*4652.68*	*4685.07*

be used to model inter–individual correlation rather than just overdispersion. To our knowledge, the first work in this direction has been produced by Mano and Suto (2014), using a Bayesian framework. None of the datasets that we have investigated does provide such hierarchical information, so we did not investigate this avenue further.

A third model class which should be mentioned here are two–part models, which, rather than allowing zeros to be generated via two different routes as in the ZIP model, define a separate model for zero– and non–zero response, where the latter part could be described by e.g. a truncated Poisson distribution (Alfò and Maruotti, 2010). Such 'Hurdle' models have the appealing property of being based on a clear hierarchical structure: first, a decision is made on whether a zero is chosen or not, and secondly, the non–zero part of the model is invoked if chosen. These models, which are beyond the scope of the present manuscript, appear promising in the context of radiation biodosimetry and so deserve further investigation.

3.5 Discussion on how to model the zero–inflation parameter

Based on the results shown in Tables 3–6, the three considered forms of modelling the zero–inflation parameter p_i provide similar results in terms of the log–likelihood. However, looking at the fitted values of this parameter, it can be observed that they can be very different depending on the specified model.

Figure 2 shows the fitted values of the parameter p_i after fitting a ZIP regression model to data (C1–C3) and (A3) (left panel) and a ZINB1 regression model to data (D1–D3) and (B2) (right panel). The solid dots represent the fitted p_i when these do not depend on covariates, and the dashed and solid lines give the fitted values when p_i is modelled through a logit link as a linear function of the dose with and without intercept, respectively.

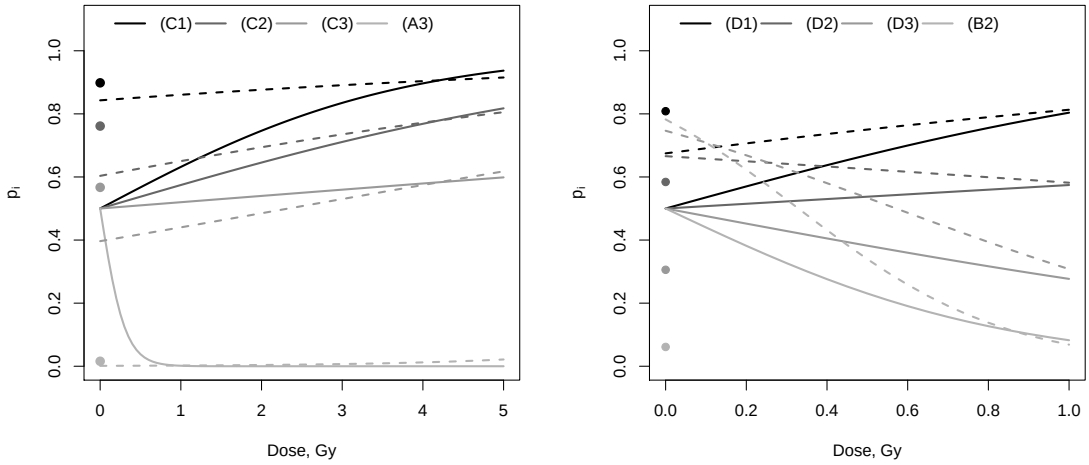

Figure 2 Fitted zero–inflation (mixture) parameters p_i as a function of dose, x_i. Solid lines correspond to modelling the mixture parameter as $\text{logit}(p_i) = \gamma_1 x_i$ and dashed lines correspond to modelling it as $\text{logit}(p_i) = \gamma_0 + \gamma_1 x_i$. Solid dots indicate the fitted probabilities when p_i is modelled as a constant, $\text{logit}(p_i) = \gamma_0$. Left panel: Results obtained from fitting a ZIP regression model to data (A3) and (C1–C3). Right panel: Results obtained from fitting a ZINB1 regression model to data (B2) and (D1–D3).

Both plots show that the mixture parameter takes similar values at the highest doses observed in each case, independently of how it is modelled. Moreover, the value of p_i is influenced by the percentage of unirradiated blood, as expected (IAEA, 2011). However, for the lowest doses, it takes very different values. If p_i is modelled as $\text{logit}(p_i) = \gamma_1 x_i$ $(i = 1, \ldots, d)$, then the mixture parameter is equal to 0.5 at zero dose. That is, the model (4) imposes the probability 0.5 of extra zeros for non–irradiation. However, this may be a very restrictive assumption. In order to allow for more flexibility, an intercept is included in model (5). If $\text{logit}(p_i) = \gamma_0 + \gamma_1 x_i$, different situations can occur. For example, the dashed lines in the right panel in Figure 2 show that for non–irradiated blood samples, the probability of extra zeros is quite similar for the four datasets (as it would be expected). But, the dashed lines in the left panel show that the probability takes very different values at dose 0. This different behavior may be explained by the first dose observed in each case. For data in the right panel, the smallest dose used was 0.1 Gy so, it is expected that the four datasets perform similarly (at dose 0, the four datasets should be practically equal). In the other hand, for data in the left panel, the smallest dose was 1 Gy and so, the value of p_i is already influenced by the percentage of irradiated blood.

Model (5) is especially meaningful for fitting (C1–C3) and (D1–D3). In a partial body exposure simulation experiment, where a fixed proportion of blood f is irradiated (for instance, 25%, 50% and 75%) to a dose x, this proportion f is not the same as the proportion $(1 - p)$ of irradiated cells in the zero–inflated model. Moreover, the magnitude of the difference also depends on the dose. The reason is that not all the irradiated cells transform and survive to metaphase, and those which do not survive can not be scored. According to Lloyd and Edwards (1983), the survival rate of the irradiated cells $s(x)$ follows a decreasing

exponential function of the dose x of the form, $s(x) = \exp(-\gamma_1 x)$. Note that for a dose of $x = 0$ the survival rate is 100%.

Suppose that in a partial body exposure we have N irradiated cells and N_0 non irradiated cells. It is clear that the proportion of irradiated blood is,

$$f = \frac{N}{N_0 + N}\,, \tag{6}$$

but the proportion of scored irradiated cells (those which survive) is,

$$(1-p) = \frac{Ns(x)}{N_0 + Ns(x)}\,. \tag{7}$$

Replacing N_0 in (7) by that isolated from (6), we obtain,

$$(1-p) = \frac{Ns(x)}{N(1-f)/f + Ns(x)} = \frac{\exp(-\gamma_1 x)}{(1-f)/f + \exp(-\gamma_1 x)} = \frac{1}{1+\exp(\gamma_0 + \gamma_1 x)}\,, \tag{8}$$

where $\mathrm{logit}(f) = -\gamma_0$. This implies the relationship $\mathrm{logit}(p) = \gamma_0 + \gamma_1 x$, justifying model (5).

The value of γ_1 depends on the kind or radiation and its capacity to damage the cells, and γ_0 is related to the fraction of irradiated blood.

Our application studies have demonstrated little difference in terms of log–likelihood between the three methods of modelling the mixture parameter. However, for partial body irradiation scenarios, we have shown that model (5) is conceptually preferable. Dataset (C2) constitutes an example where this conceptual advantage led to a superior practical performance.

4 Simulation study

In this section, we will give some more objective evidence for our claim that overdispersion and zero–inflation are in general separately identifiable. If that is true, we would expect

- the ZINB model to be favorable if both of these features are present;
- the NB model to be favorable if only overdispersion is present;
- the ZIP model to be favorable if only zero–inflation is present;
- the Poisson model to be favorable if none of these features are present.

Therefore, we generated 100 data sets from each of Poisson, ZIP, NB2 and ZINB2 models, then we fitted the data using all four models, and counted the proportion of times that each model gives the 'winning result' in terms of AIC and BIC. We also computed the score tests introduced earlier (where applicable) and give the proportions of rejection of the respective null hypothesis. For the data generation, we used the Poisson model fitted to (A3) as base model (we know from our previous analysis that this is a 'correct' model), with five doses $x_1 = 1, \ldots, x_5 = 5$. Then we instilled successively zero–inflation and overdispersion into this data–generating process, and observed the outcome. For the zero–inflation parameter p_i, we assumed scenario (3); that is, we did not assume dependence of this parameter on dose.

For the simulation of the ZIP models, we distinguished between (a) mild zero–inflation [$p = 0.1$], (b) moderate zero–inflation [$p = 0.2$] and (c) strong zero–inflation [$p = 0.5$]. For the NB models, we tried to match the degree of 'non–Poissonness" according to the following reasoning. Note that, for ZIP models, one has

$$\mathrm{Var}(Y_{ij}) = \mu_i(1 + p\lambda_i) = \mu_i\left(1 + \frac{p}{1-p}\mu_i\right).$$

Table 7 Proportion of correctly identified models using AIC, for models using the identity–link (top) and log–link (bottom). The 'correct' model choice is provided in bold letters. Columns add up to 100%.

	True model	P	ZIP			NB			ZINB		
link	Fitted model		mild	mod.	strong	mild	mod.	strong	mild	mod.	strong
	P	**91**	0	0	0	0	0	0	0	0	0
id	ZIP	6	**88**	**96**	**96**	5	0	0	8	0	0
	NB	3	0	0	0	**92**	**91**	**90**	3	0	2
	ZINB	0	12	4	4	3	9	10	**89**	**100**	**98**
	P	**95**	0	0	0	2	0	0	0	0	0
log	ZIP	3	**50**	**67**	**96**	42	41	7	9	0	0
	NB	2	0	0	0	**51**	**54**	**85**	7	1	1
	ZINB	0	50	33	4	5	5	8	**84**	**99**	**99**

Table 8 Proportion of correctly identified models using BIC, for models using the identity–link. The 'correct' model choice is provided in bold letters. Columns add up to 100%.

	True model	P	ZIP			NB			ZINB		
link	Fitted model		mild	mod.	strong	mild	mod.	strong	mild	mod.	strong
	P	**100**	0	0	0	13	0	0	0	0	0
id	ZIP	0	**95**	**100**	**100**	6	0	0	45	1	0
	NB	0	2	0	0	**81**	**100**	**100**	48	25	17
	ZINB	0	3	0	0	0	0	0	**7**	**74**	**83**
	P	**100**	1	0	0	23	0	0	0	0	0
log	ZIP	0	**52**	**68**	**100**	22	41	7	51	1	0
	NB	0	11	0	0	**55**	**59**	**93**	39	14	15
	ZINB	0	36	32	0	0	0	0	**10**	**85**	**85**

For the negative binomial model (NB2), we know that

$$\text{Var}(Y_{ij}) = \mu_i(1 + \alpha\mu_i);$$

hence, for an equal degree of non-Poissonness we can equate $\alpha = p/(1-p)$. Following this reasoning, we considered in our simulation study data generated from a NB2 distribution with parameters (a) $\alpha = 1/9$ (mild overdispersion), (b) $\alpha = 1/4$ (moderate overdispersion) and (c) $\alpha = 1$ (strong overdispersion). For the ZINB models, we considered the pairings (a) mild/mild, (b) moderate/moderate and (c) strong/strong.

Tables 7 and 8 indicate clearly that, in the vast majority of cases, the underlying models were correctly identified. For the log–link, we observed a tendency of mildly zero–inflated Poisson models to be classified as ZINB models, and mildly overdispersed (NB) models to be classified as ZIP models. The stronger the overdispersion or zero–inflation, the better are the associated models separately identifiable. The proportion of correctly identified models was generally larger for the identity– than for the log–link, and was generally larger when using BIC rather than AIC. The only exception to this are 'mild/mild' zero–inflated negative binomial models, which tend to be classified as ZIP or NB models under BIC. The score tests in Table 9 speak a very clear language: The proportion of rejection of the Poisson and ZIP model is close to 0, when these models are true, and close or equal to 1, when these are false. Overall these simulations confirm impressively the separate identifiability of zero–inflated and overdispersed models, as well as the need for models which are both overdispersed and zero–inflated.

Table 9 Proportion of rejection of the smaller model using score tests (at the 5% level of significance), for models using the identity link (top row) and the log–link (bottom three rows). Only values in bold are fully meaningful as in this case the true model corresponds to one of the two models tested against.

	True model	P	ZIP			NB			ZINB		
link	Fitted model		mild	mod.	strong	mild	mod.	strong	mild	mod.	strong
id	P/ZIP	**0.05**	**1**	**1**	**1**	0.86	1	1	1	1	1
	P/ZIP	**0.02**	**1**	**1**	**1**	0.90	1	1	1	1	1
log	P/NB	**0.03**	0.99	1	1	**0.98**	**1**	**1**	1	1	1
	ZIP/ZINB	0.00	**0.03**	**0.01**	**0.05**	0.84	1	1	**0.77**	**1**	**1**

Table 10 Summary of recommended settings under different exposure scenarios, when counting dicentrics and centric rings (low LET and high LET correspond to sparsely and densely ionising radiation, respectively). When analysing micronuclei, we would advocate the use of ZINB models irrespective of the exposure pattern.

exposure		whole body	partial
LET	low	Poisson/NB	ZIP
	high	NB/Neyman A	ZINB

5 Concluding remarks

Zero–inflated models have been proposed for modelling the number of aberrations per cell as a function of the dose. They have been compared with other models showing that they behave well in several scenarios, especially for partial body exposure. Moreover, results obtained by modelling the mean yield of aberrations through a log–link were compared with those ones obtained by using the identity link showing that both link functions give very similar results. Score tests justified the use of zero–inflated models for fitting several datasets. For the problem of testing a Poisson versus a ZIP model, we have presented in this manuscript a variant of van den Broek's score test which allows for the use of the identity link.

A relevant finding of this paper is that overdispersion needs to be taken into account irrespective of whether the data stem from full or partial body exposure. In the case of full body exposure, for densely ionising radiation or when micronuclei are analysed, the overdispersion will be relatively high and can often be addressed through a (possibly zero–inflated) negative binomial model or the Neyman A model, whereas for sparsely ionising radiation overdispersion will be relatively mild (but not always ignorable) and can often be addressed exchangeably through a negative binomial or a zero–inflated Poisson model (or even other models). Partial body exposure will in general require explicit modelling of the zero-inflation. While for sparsely ionising radiation zero–inflated Poisson models turned out to be sufficient in our analysis, for densely ionising radiation it was generally necessary to model the overdispersion *on top of* the zero–inflation, through a zero–inflated negative binomial model. A small simulation study has confirmed that the concept of considering overdispersion and zero–inflation as separately identifiable model properties is sensible.

Table 10 summarizes our recommended settings for different exposure scenarios. The important message from this table is that models which allow for overdispersion will be needed in the bottom row (due to the densely ionising radiation), and that zero-inflated models will need to be used in the right column (where the body exposure is only partial). The table should not be considered in an 'exclusive' sense – there will often be many other models which will fit well too. We have chosen the named models based on conceptual plausibility, and practical performance in our analysis in Tables 3 to 6. If one is in doubt about the exposure scenario, ZIP models (especially those which model the zero–inflation parameter linearly) will generally lead to good results.

Zero–inflated models are also directly biologically relevant, as partial body exposures always lead to a mixture of non–irradiated and irradiated blood lymphocytes within the body at the time of irradiation, and blood sampling for biological dosimetry takes place >24 hours after exposure, the timescale for full circulation of lymphocytes within the human body, so the exposed and un–exposed fractions can reasonably be expected to be fully mixed within the sample taken.

One issue that we have not discussed in this paper is how, given a fitted model, the dose can be estimated from the fitted model for a given aberration count. This is an inverse regression problem; two Bayesian–like solutions to which have been recently provided by Higueras et al. (2015a, 2015b) in whole and partial body exposure scenarios, respectively. Higueras et al. (2015a) can effectively be used for Poisson, NB1, Neyman A and Hermite ($r = 2$), and can be extended for all two parameter compound Poisson count distributions, this includes Poisson–inverse Gaussian and Pólya–Aeppli models. The approach by Higueras et al. (2015b) can be used for the ZIP(3) distribution. A well–fitting model is, however, absolutely crucial for the success of these techniques. We hope that our manuscript could contribute to addressing this question.

The models used with the cytogenetic example data presented in his work would certainly be applicable to other fields, specifically including nuclear radiation and technology research where the Poisson distribution is frequently applied but also potentially for chemical and other mutagens. Indeed the results of this work have shown that it is useful to formally assess the most appropriate models in a dynamic way wherever count data appear, or models are used to formally assess effects on biological systems.

Acknowledgements
This report is independent research supported by the National Institute for Health Research, Research Methods Opportunity Funding Scheme entitled "Random effects modelling for radiation biodosimetry" (NIHR-RMOFS-2013-03-4). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. This work has also been partially supported by the grant MTM2013-41383P from the Spanish Ministry of Economy and Competitiveness co-funded by the European Regional Development Fund (EDRF). The authors wish to thank the Editors and three anonymous referees for their thoughtful and constructive comments.

Conflict of Interest
The authors have declared no conflict of interest.

Appendix

A.1. Poisson against zero–inflated Poisson: Score statistic under identity link

We adapt van den Broek's (1995) score test, which uses the log–link, for the use of the identity link. Assume that $p_i \equiv p$ is constant across observations. We denote $\theta = p/(1-p)$, then testing the null hypothesis $H_0 : p = 0$ against $H_1 : p \neq 0$ is equivalent to testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. The likelihood of the ZIP model is given by

$$L_{ZIP}(\theta, \boldsymbol{\beta}) = \frac{1}{(1+\theta)^n} \prod_{i=1}^{n} \left\{ \mathbf{1}_{\{y_i=0\}}(\theta + e^{-\lambda_i}) + \mathbf{1}_{\{y_i \neq 0\}} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right\},$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Taking first derivatives of $l_{ZIP}(\theta, \boldsymbol{\beta}) = \log L_{ZIP}(\theta, \boldsymbol{\beta})$ under the identity link $\lambda_i = \mathbf{x}_i^T \boldsymbol{\beta}$, one obtains

$$\frac{\partial l_{ZIP}}{\partial \theta} = \sum_{i=1}^{n} \left\{ \frac{-1}{1+\theta} + \mathbf{1}_{\{y_i=0\}} \left(\frac{1}{\theta + \exp(-\lambda_i)} \right) \right\} \tag{9}$$

$$\frac{\partial l_{ZIP}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i \left\{ \mathbf{1}_{\{y_i=0\}} \left(\frac{-e^{-\lambda_i}}{\theta + e^{-\lambda_i}} \right) + \mathbf{1}_{\{y_i>0\}} \left(\frac{y_i}{\lambda_i} - 1 \right) \right\} \tag{10}$$

which under $H_0 : \theta = 0$ gives the score vector

$$S(0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\mathbf{1}_{\{y_i=0\}} e^{\lambda_i} - 1, \mathbf{x}_i \left(\frac{y_i}{\lambda_i} - 1 \right) \right). \tag{11}$$

Note that the right hand side of this is just the score-statistic of a Poisson-GLM under the identity link. Hence, at the ML estimate under H_0 this term would be 0. However, due to the constraints that need to be employed to keep the Poisson mean positive, this expression will usually not be exactly 0. Hence, the full expression (11) will be used for the test statistic. Taking once more partial derivatives of (9) and (10), one obtains the (expected) Fisher information $J(0, \boldsymbol{\beta})$ under $H_0 : \theta = 0$ as

$$J(0, \boldsymbol{\beta}) = -E \left(\begin{array}{cc} \frac{\partial^2}{\partial \theta^2} & \frac{\partial^2}{\partial \theta \partial \boldsymbol{\beta}^T} \\ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \theta} & \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \end{array} \right) \Bigg|_{\theta=0} = \sum_{i=1}^{n} \left(\begin{array}{cc} e^{\lambda_i} - 1 & -\mathbf{x}_i^T \\ -\mathbf{x}_i & \frac{1}{\lambda_i} \mathbf{x}_i \mathbf{x}_i^T \end{array} \right)$$

The test statistic is given by

$$T = S(0, \hat{\boldsymbol{\beta}})^T J(0, \hat{\boldsymbol{\beta}})^{-1} S(0, \hat{\boldsymbol{\beta}})$$

with $\hat{\lambda}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ estimated under the Poisson model. In all experiments which we have carried out we found excellent adherence of the distribution of T to the $\chi^2(1)$ distribution under H_0. For instance, for 1000 data sets generated from the fitted Poisson model (A3), Figure 3 (left) gives the distribution of T with the $\chi^2(1)$ distribution overlaid, and Figure 3 (right) gives the power in comparison to the log-link version of the test. We see that both tests behave well and similarly, with strong test powers and good attainment of the nominal significance level.

Form a theoretical point of view, the $\chi^2(1)$ property may appear surprising since the logit link (3) employed for the mixture parameter p induces the constraint $p \geq 0$. However, since the score test does not use the model fit under the alternative (so, here, under the ZIP model), it is not affected by this constraint, so that the $\chi^2(1)$ distribution remains intact, in accordance with van den Broek (1995) and usual likelihood theory. In order to obtain a genuinely one–sided test (that is, a test which alerts at zero–inflation but not at deflation), one needs to construct an adjusted version of the score–test as outlined in a more general setting in Molenberghs and Verbeke (2007), noting computational challenges. This one–sided version would then follow a $0.5[\chi^2(0) + \chi^2(1)]$ distribution. For comparability and simplicity, we have used all tests employed in this manuscript in their (implicitly) two–sided version, with critical values taken from the $\chi^2(1)$ distribution.

References

Alfò, M., and Maruotti, A. (2010). Two-part regression models for longitudinal zero-inflated count data. *The Canadian Journal of Statistics*, **38**, 197–216.

Ainsbury, E.A., Vinnikov, V.A., Maznyk, N.A., Lloyd, D.C. and Rothkamm, K. (2013). A Comparison of six statistical distributions for analysis of chromosome aberration data for radiation biodosimetry. *Radiation Protection Dosimetry*, **155**, 253–267.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Böhning D., Dietz E., Schlattmann P., Mendoca L. and Kirchner U. (1999). The zero–inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society A*, **162**, 195–209.

Brame, R.S. and Groer, P.G. (2002). Bayesian analysis of overdispersed chromosome aberration data with the negative binomial model. *Radiation Protection Dosimetry*, **102**, 115–119.

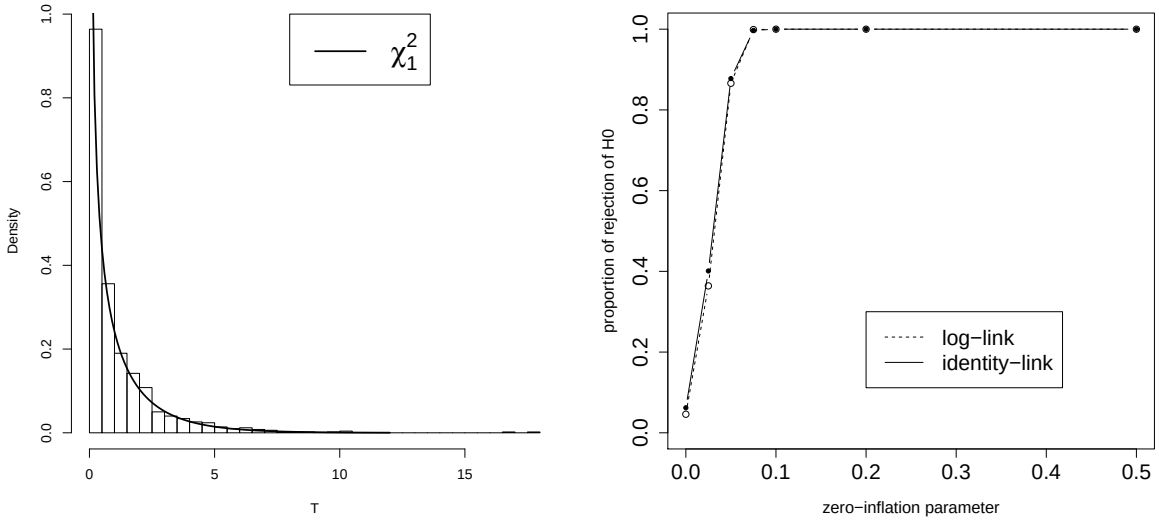

Figure 3 Left: Null distribution of T for 1000 data sets generated from the Poisson model fitted to data set (A3). Right: Power of the introduced test in comparison to the log–link version, for each 1000 data sets generated by zero–inflating the fitted Poisson model (A3) with ZIP parameter $p = 0$, 0.025, 0.05, 0.075, 0.1, 0.2 and 0.5.

Cheung Y.B. (2002). Zero–inflated models for regression analysis of count data: a study of growth and development. *Statistics in Medicine*, **21**, 1461–1469.

Dean, C.B. and Lawless, J.F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, **84**, 467–472.

Di Giorgio, M., Edwards, A. A., Moquet, J. E., Finnon, P., Hone, P. A., Lloyd, D. C., and Valda, A. (2004). Chromosome aberrations induced in human lymphocytes by heavy charged particles in track segment mode. *Radiation protection dosimetry*, **108**, 47–53.

Greene, W.H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Papers from New York University, Leonard N. Stern School of Business, Department of Economics.

Gudowska–Nowak, E., Lee, R. Nasonova, E., Ritter, S. and Scholz, M. (2007). Effect of LET and track structure on the statistical distribution of chromosome aberrations. *Advances in Space Research*, **39**, 1070–1075.

Hall,, D. B. (2000). Zeroinflated Poisson and binomial regression with random effects: a case study. *Biometrics*, **56**, 1030-1039.

Hall, E. J., Giaccia, A. J. (2012). *Radiobiology for the radiologist*, 7^{th} edition. Lippincott Williams & Wilkins. Philadelphia.

Hlatky, L., Farber, D., Sachs, R. K., Vazquez, M., Cornforth, M. N. (2002). Radiation–induced chromosome aberrations: insights gained from biophysical modelling. *BioEssays*, **24**, 714–723.

Heimers, A., Brede, H.J., Giesen, U. and Hoffmann, W. (2006). Chromosome aberration analysis and the influence of mitotic delay after simulated partial–body exposure with high doses of sparsely and densely ionising radiation. *Radiation and Environmental Biophysics*, **45**, 45–54.

Higueras, M., Puig, P., Ainsbury, E.A., Rothkamm, K. (2015a). A new inverse regression model applied to radiation biodosimetry. *Proceedings of the Royal Society A*, DOI: 10.1098/rspa.2014.0588.

Higueras, M., Puig, P., Ainsbury, E.A., Vinnikov, V.A, and Rothkamm, K. (2015b). A new Bayesian model applied to cytogenetic partial body irradiation estimation. *Radiation Protection Dosimetry*, DOI: 10.1093/rpd/ncv356.

IAEA (2011). *Cytogenetic Dosimetry: Applications in Preparedness for a Response to Radiation Emergencies*. International Atomic Energy Agency. Vienna.

Lambert, D. (1992). Zero–inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Lee A.H., Wang K. and Yau K.K.W. (2001). Analysis of zero–inflated Poisson data incorporating extent of exposure. *Biometrical Journal*, **43**, 963–975.

Lloyd, D. and Edwards, A. (1983). Chromosome aberrations in human lymphocytes: effect of radiation quality, dose, and dose rate. In *Radiation-Induced Chromosome Damage in Man* (T. Ishihara and and M. Sasaki, Eds.), pp. 2349. Alan R. Liss, New York, 1983.

Mano, S. and Suto, Y. (2014). A Bayesian hierarchical method to account for random effects in cytogenetic dosimetry based on calibration curves. *Radiat Environ Biophys.* **5**3, 775–780.

Molenberghs, G. and Verbeke, G. (2007). Likelihood Ratio, Score, and Wald Tests in Constrained Parameter Space. *The American Statistician, **61**, 22–27.*

Puig, P. and Valero, J. (2006). Count data distributions: some characterizations with applications. *Journal of the American Statistical Association*, **101**, 332–340.

Puig, P. and Barquinero, J. (2011). An application of compound Poisson modelling to biological dosimetry. *Proceedings of the Royal Society A*, **467**, 897–910.

Ridout, M., Hinde, J. and Demétrio, C. G. (2001). A score test for testing a zero–inflated Poisson regression model against zero–inflated negative binomial alternatives. *Biometrics*, **57**, 219–223.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R., Chakravarti, I. M. (1956). Some small sample tests of significance for a Poisson distribution. *Biometrics*, **12**, 264–282.

Romm, H., Ainsbury, E., Barnard, S., Barrios, L., Barquinero, J. F., Beinke, C., Deperas, M., Gregoire, E., Koivistoinen, A., Lindholm, C., Moquet, J., Oestreicher, U., Puig, P., Rothkamm, K., Sommer, S., Thierens, H., Vandersickel, V., Vral, A., Wojcik A. (2013). Automatic scoring of dicentric chromosomes as a tool in large scale radiation accidents. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, **756**, 174–183.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, **51**, 738–743.

Virsik, R.P. and Harder, D. (1981). Statistical interpretation of the overdispersed distribution of radiation–induced dicentric chromosome aberrations at high LET. *Radiation Research*, **85**, 13–23.

