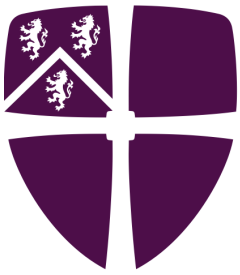
# Estimating dose and exposure fraction from radiation biomarkers in the presence of overdispersion



Adam Errington

Supervisors: Dr. Jochen Einbeck
Dr. Jonathan Cumming

Department of Mathematical Sciences
Durham University

*Doctor of Philosophy*

November 2022

# Declaration

I hereby declare that except where a specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.



Adam Errington

November 2022

# Acknowledgements

I would like to give special mentions and thank yous to my supervisors, Jochen Einbeck and Jonathan Cumming, with whom I have had the pleasure to work with during this project. I am very grateful for their advice and support which have been invaluable in the completion of this thesis. I would also like to thank the Cytogenetics Group at Public Health England, Didcot, UK, for providing the PHE-Foci2 $\gamma$-H2AX data. Finally, I would like to thank my family, whose love and guidance are always with me in whatever I pursue.

# Abstract

It is typically assumed that the total $\gamma$-H2AX foci produced in a sample of blood cells is Poisson distributed, whose expected yield can be represented by a linear function of the absorbed dose. However, in practice, because of unobserved heterogeneity in the cell population, the standard Poisson assumption of equidispersion will most likely be contravened which will cause the variance of the foci counts to be larger than their mean. In both whole and partial body exposure this phenomenon is perceptible, unlike in the context of the dicentric assay in which overdispersion is usually considered only to be linked to partial exposure. For such situations, and as we will demonstrate, it is suitable to utilise a model that can handle overdispersion such as the quasi-Poisson or negative binomial regression.

The scenarios of most radiation accidents result in partial-body exposures or non-uniform dose distribution, leading to a differential exposure of lymphocytes in the body. Subsequently for the exposed individuals, their blood will contain a mixture of cells showing no radiation impact at all and cells featuring a distribution of counts according to dose of exposure. For such exposure scenarios, it remains that there are no statistical procedures to follow for the $\gamma$-H2AX assay. Part of this work will focus on updating the contaminated Poisson method, traditionally used in conjunction with cytogenetic biomarkers, to enable an estimate of the radiation dose and irradiated fraction to be found in the presence of both zero-inflation and overdispersion. As an extension, we discuss and compare how to measure the uncertainty associated with a given dose estimate via the delta, Merkle and ISO methods. We illustrate their applications firstly via simulated zero-inflated Poisson and NB1 data, with the non-inflated part being generated using an external $\gamma$-H2AX whole-body calibration curve, before applying the methodology to practical data.

# Table of contents

# List of symbols

| | |
|---|---|
| $y_{ij}$ | $j$-th raw data count of the $i$-th slide |
| $n_i$ | Total observations of the $i$-th slide |
| $y_i$ | Yield of the $i$-th slide |
| $k$ | Total number of slides |
| $E(.)$ | Expectation operator |
| $x_i$ | Predictor variable for $i$-th dose |
| $n^*$ | Total observations per single sample |
| $\sum$ | Summation operator |
| $s^*$ | Sum of observations per single sample |
| $y^*$ | Yield per single sample |
| $\beta$ | Vector of regression parameters |
| $\hat{\beta}$ | Vector of regression coefficient estimates |
| $x^*$ | Dose estimate from single sample |
| $SE(.)$ | Standard error |
| $\partial$ | Differential operator |
| $\chi^2$ | Chi-square statistic |
| $P(.)$ | Probability density function |
| $\int$ | Integral operator |
| $\propto$ | Proportional to |
| $U(.)$ | Uniform prior |
| $\phi$ | Dispersion |
| $Var(.)$ | Variance operator |
| $\hat{\delta}$ | Index of dispersion |
| $N$ | Total number of counts over $k$ slides |
| $\bar{y}$ | Yield over $k$ slides |
| $U$ | Papworth test statistic |
| $S$ | Sum of observations over $k$ slides |
| $l$ | Dummy subscript |
| $h$ | Dummy subscript |

| | |
|---|---|
| $N_0$ | Total number of zeros (aberration-free cells) over k slides |
| $n_0$ | Total number of zeros per slide |
| $\Phi(.)$ | Cumulative density function |
| $zi$ | Zero-inflation index |
| $g(.)$ | Canonical link function |
| $\mu$ | Distribution mean |
| $\lambda$ | Mean yield (via the calibration curve) |
| $L(.)$ | Model likelihood |
| $\rho$ | Constant used to represent degree of exposure |
| $X^2$ | Pearson goodness-of-fit statistic |
| $\nu$ | Residual degrees of freedom |
| $w$ | Number of model parameters |
| $\alpha$ | Overdispersion parameter |
| $c$ | Index to identify form of NB/ZINB distribution |
| $\Gamma(.)$ | Gamma function |
| $m$ | Dummy subscript |
| $p$ | Zero-inflation parameter |
| $\ell(.)$ | Model log-likelihood |
| $I(.)$ | Indicator function |
| $\gamma_0$ | Constant to infer type of radiation and cell damage |
| $\gamma_1$ | Constant to infer proprtion of irradiated blood |
| $SD(.)$ | Standard deviation |
| $t_W$ | Welch test statistic |
| $t_M$ | Mann-Whitney/Wilcoxen test statistic |
| $\kappa(.)$ | Dose-dependent function describing the survival rate of irradiated cells |
| $D$ | Constant for absorbed dose |
| $R$ | Likelihood ratio test statistic |
| $W$ | Wald test statistic |
| $\epsilon$ | Level of significance for hypothesis testing |
| $\pi_0$ | Probability of a zero observation |
| $p_0$ | Total proportion of zeros over $N$ cells |
| $Sc(.)$ | Score function |
| $det(.)$ | Matrix determinant |
| $F$ | Irradiated fraction |
| $q$ | Probability constant |
| $u$ | Random effect |
| $\sigma_r^2$ | Random effect variance |
| $\sigma_\epsilon^2$ | Residual variance |

| | |
|---|---|
| $Z$ | Random indicator variable |
| $\tau$ | Constant for length of correlated strings |

# List of figures

# List of tables

# Chapter 1

# Introduction

Today, there remains a crucial shortage of methods capable of determining the extent of exposures of human beings to ionising radiation (IR). However, knowledge of individual exposures is essential for early triage during radiological incidents to provide optimum medical procedures to each person. Members of the public will usually not be carrying a personal dosimeter, thus other procedures to rapidly and reliably determine the level of exposure and contracted dose are required.

The main purpose of biological dosimetry is to assess the amount of induced radiation damage at a cellular level. The quantification of the radiation dose absorbed is particularly useful to distinguish those deemed to be "critically exposed", who should be prioritised, from the "worried well", people who have (comparatively) been minimally exposed and unlikely to need urgent treatment [74]. As well as this, it can provide us with useful information regarding probable future health consequences, both stochastic and deterministic, for victims of radiation incidents [101]. Of course, it is natural to continue developing novel and more powerful medical precautions, therefore necessitating in the discovery of new biomarkers [61, 19, 50] but also in efforts to explore the maximum potential of existing biomarkers.

## 1.1 Role of radiation biomarkers

Current diagnostics are based on radiation biodosimetry, a field that has seen enormous progress within the last decade. Considerable effort has been put into the development of radiation exposure biomarkers, which would provide information about the effective radiation dose [9, 79]. The biodosimetric system is capable of identifying radiation exposure by application in mass screening settings [100, 99].

The ideal radiation biomarker (or dosimeter) should provide information about dose and time and should be independent of environmental and confounding factors such

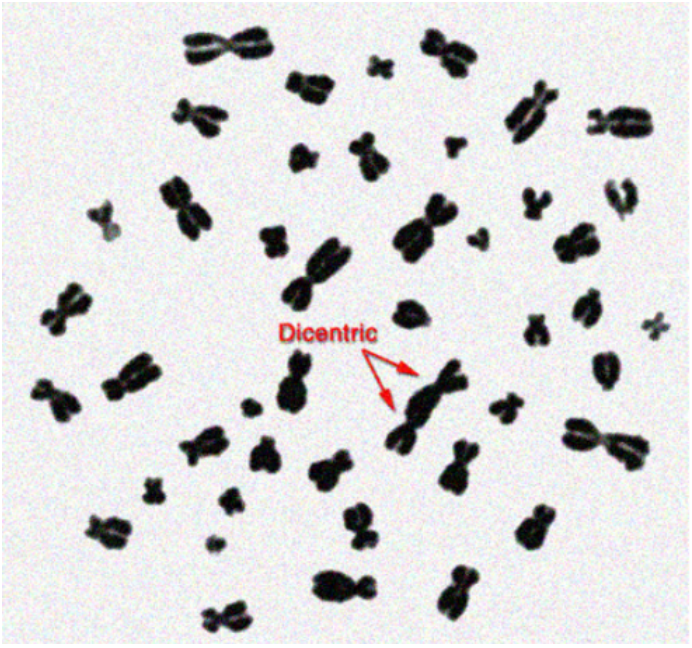



Fig. 1.1 Red arrows indicating the two centromeres on the dicentric chromosome [27].

as smoking, drug therapy, age, etc. [80, 55, 69]. Such a biomarker obviously does not exist, however for suitability in triage it is desirable for a biomarker to possess (at least) some of the following qualities; present low background, low donor variability, ease of sampling, low cost, rapid analysis and work as a "risk-marker". Although biomarkers based on gene-expression and micronuclei have proven to be competitive and gained popularity in recent years [105, 78], in this thesis we will dedicate our attention to the $\gamma$-H2AX biomarker, and to a limited extent, the dicentric biomarker.

### 1.1.1 The dicentric assay

Depending on the dose and the amount of exposure to IR, there may be significant consequences for the victims. The main target of IR is DNA, which can be damaged indirectly through reactive oxygen species (ROS) or directly through double-strand breaks [51]. Due to their lack of inter-individual variation [29], the "gold standard" method for the detection of double strand breaks (DSBs) is based on the scoring of dicentric chromosomes. Specifically, this is a chromosome with two centromeres (“crossings”) instead of the usual one as shown in Fig 1.1. It is formed by an exchange between the centromeric pieces of two broken chromosomes, and in the complete form the resultant dicentric chromosome is accompanied by an acentric fragment which is composed of the remaining pieces of the broken chromosomes and does not contain a centromere [53].

It was first proposed by Bender and Gooch [13] that the dicentric counts observed in metaphases from peripheral lymphocytes could be used for dose evaluation of

human radiation exposures. A further significant development of the method was the introduction of the fluorescence plus Giemsa (FPG) staining [81]. With this process it became possible to distinguish between the first and following mitotic divisions after culture initiation, which is important as dicentric chromosomes are lost at cell division.

For decades, the analysis of dicentrics has been considered to be the most reliable cytogenetic endpoint for biological dosimetry, since dicentrics are easily scorable and the control level is low [53]. Due to the amount of research that has been conducted and the widespread utilisation of this assay, the dicentric chromosome should be considered to be a "best possible" albeit imperfect choice of biomarker, as it has a few primary limitations. Firstly, reliable samples cannot be taken immediately after exposure as it takes at least 2 days to obtain suitable metaphase spreads following irradiation and subsequent stimulation of lymphocytes [94]. The analysis itself is both time-consuming and requires experienced cytogeneticists in order to produce an accurate assessment of the level of radiation damage. As a result, the total number of cases that can be assessed globally in any given week is approximately 3000 [94]. This means in a large-scale radiation incident, triage of casualties may well be dangerously slow, potentially posing long-term harm to victims' health. Therefore, other biomarkers should be investigated which allow faster assessment.

### 1.1.2 Protein-based biomarkers

Chromosomes in an organism are made of a substance called chromatin, which itself consists of nucleosomes in more complex, higher order structures. These nucleosomes are composed of both DNA and octamers of histones, groups of eight proteins that are used to package the DNA double helix. Specifically, DNA is wrapped around the eight protein structure. Each octamer is made from four types of histone, H2A, H2B, H3, and H4, and each type of histone is represented twice [89]. The H2A histone has four subtypes, which are grouped into three subfamilies: H2A1-H2A2, H2AZ, and H2AX. The focus of this thesis will be on H2AX, which can account for anything from 2% to 20% of the H2A histones found in human cells.

The H2AX histone is a DNA-repair protein; that is, once a cell gets exposed to ionising radiation and a DSB has occurred, it coordinates the repair of the damaged DNA and in this process phosphorylates, becoming $\gamma$-H2AX [89]. This phosphorylation leads, after addition of fluorophore-labelled antibodies, to fluorescent dots which can be counted under a microscope. The phosphorylation is only visible for up to approximately 24 hours after radiation exposure. DSBs are spontaneously induced at a very low rate, and very few other biologically relevant processes induce them, so the presence of a significant amount of DSBs implies that an organism has been exposed to ionising radiation.

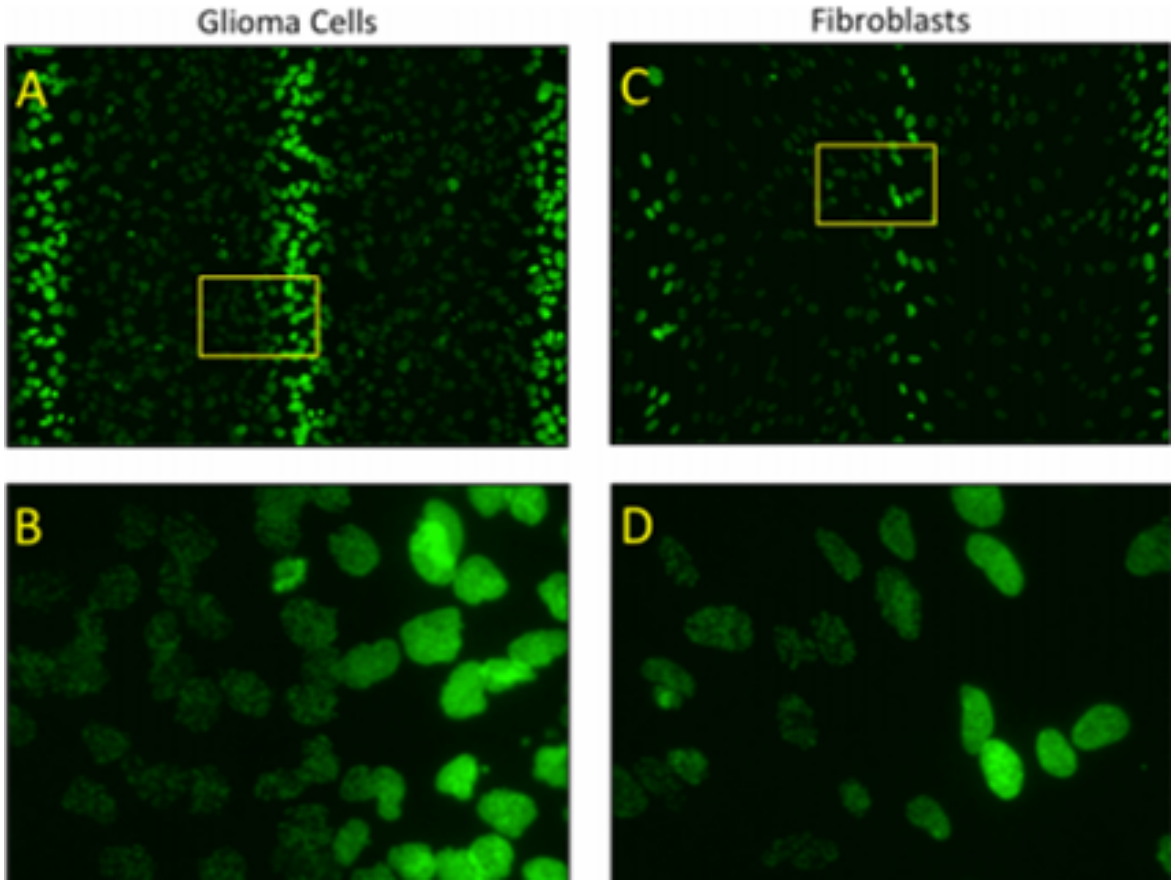



Fig. 1.2 $\gamma$-H2AX immuno-stained cells following irradiation by an array of synchrotron microbeams at a dose of 283 Gy. DNA double strand breaks are visualized via $\gamma$-H2AX staining in cultured human glioma cells (A) and human fibroblasts (C). Boxed regions are shown in (B) and (D) highlighting the difference in the number of $\gamma$-H2AX foci between the peak and valley regions for both cell types. [10]

Background levels of $\gamma$-H2AX foci vary between cell types. In peripheral blood lymphocytes, reported base levels are between 0.05 and 0.5 foci per cell. As $\gamma$-H2AX formation depends on an enzymatic process, analysis should take place approx 1 hour after a full body exposure, as at this point the vast majority of the induced foci are both present and at a size and intensity where they can be scored reliably. For a triage situation with a mass number of potential casualties, samples should be taken as soon as possible, ideally within 24 hours, with a potential upper limit for feasibility of approximately 3 days post-exposure [90]. In addition, the initial kinetics of foci formation and loss are inconsistent [57]. The main problem with quantifying early foci induction is that the intensity and size of each focus grow over time. The detection of foci is affected by a range of parameters, including the cell type (as illustrated in Fig 1.2) and the individual scoring criteria for manual as well as automated scoring [92]. This means, for the same biological sample, variation in foci scores is more prevalent at earlier timepoints after exposure.

The suitability of this histone as a biomarker for DSBs [60, 11], and by extension, ionising radiation exposure [70, 88, 92, 93], has long been established in the literature. However, statistical work to quantify this relationship and facilitate the actual dose estimation has only been carried out quite recently [6, 3, 28]. It should be noted that $\gamma$-H2AX foci data is not only used for biological dosimetry but much more prominently for several research questions in radiation biology [58, 56, 85, 40]. The $\gamma$-H2AX assay has also been used to quantify DNA damage induced in peripheral blood lymphocytes during diagnostic CT examinations [12, 39, 91] as well as radiotherapy [63, 95].

However, while studies reported the use of $\gamma$-H2AX foci induction following exposure to therapeutic doses of ionising radiation [68, 59], how the assay would perform at higher doses, particularly in humans, remained unclear for a while. In a study using nonhuman primates subjected to total-body irradiation in the non-lethal to lethal dose ranges [86], the authors showed that $\gamma$-H2AX analysis in lymphocytes and plucked hair follicles (eyebrows and whiskers) may be useful for estimation of radiation dose at times at least 4-days post-exposure at doses of 3.5Gy and above. In addition, the development of the RABIT (Rapid Automated Biodosimetry Tool), to respond to major radiological accidents, enabled throughput $\gamma$-H2AX analysis for radiation biodosimetry of up to 30,000 blood samples a day. Its purpose was to fully automate the $\gamma$-H2AX assay, from the isolation of human blood lymphocytes to the immunolabeling of $\gamma$-H2AX and image acquisition [37].

Given blood samples provided from an exposed patient, foci will be scored either manually in a process called immunofluorescence microscopy, or automatically through flow cytometry using machinery such as MetaCyte [16]. However, automated scoring can have consistency issues [93], so the data we are looking at has been manually scored. The main control issues to consider when scoring samples are the point at which dim foci are classified as background noise, due to their low intensity or small size, and the potential for groups of foci in close proximity to each other that can easily be perceived as fewer in number than they actually are. Both of these may be affected by differences in the optical resolution and light efficiency of the microscope and camera used for the imaging of the foci, as well as the discretion of the individual scorer. The H2AX datasets we will be using in this thesis are from 2 separate scorers (scored under the same conditions), however we identified differences in their observed foci; one recorded consistently higher foci counts than the other.

### 1.1.3 Motivation for H2AX as a DNA damage sensor

When comparing the $\gamma$-H2AX histone biomarker with its dicentric counterpart, we must first be aware that they are two different types of biomarker. The dicentric biomarker is very well established in the literature, with clear and comprehensive statistical methods, whilst the $\gamma$-H2AX histone is not. A key strength of the dicentric biomarker is that it

has very little inter-individual variation. This is not true for the $\gamma$-H2AX biomarker, which has, as well as potential inter-individual variation, far stronger inter-laboratory variation [28]. However, the time required between sampling and analysis is far shorter for $\gamma$-H2AX foci (a few hours) than for the dicentric biomarker (2-3 days). The dicentric biomarker also has a lower throughput than the $\gamma$-H2AX histone and is more labour intensive, requiring experienced and skilled cytogeneticists. As a result of this the global weekly capacity for analysis of the dicentric biomarker in "triage mode" (scoring only 50 cells per sample and with a detection limit of 0.5Gy) is approximately 3000 samples [94], clearly not practical for a situation with a high number of potential casualties. There is no currently stated upper limit for the number of $\gamma$-H2AX histone samples scored per week, but it is reasonable to assume that any such limit would be far larger than 3000.

The $\gamma$-H2AX histone biomarker operates within comparatively strict time limits: the phosphorylated foci initially form within minutes of exposure, but are typically only visible until approximately 24 hours after [35]. In comparison, while a blood sample can be taken within a few hours of a whole body exposure for the dicentric biomarker, delaying taking a sample until over 24h later is "advisable" if a non-uniform or partial body exposure is suspected. Otherwise, IAEA guidelines suggest that blood samples for analysis of this biomarker be obtained "promptly" but give no strict upper limit, suggesting that aberration yields will drop after four weeks, increasing uncertainty [53].

## 1.2 Presenting dicentric and H2AX data

The dicentric dataset, which hereafter we refer to as "**PHE-Dicentric**", consists of blood samples irradiated with three doses; 0.5, 0.7 and 1Gy of 2.1 MeV neutrons. The proportion of irradiated blood is 100% (whole-body irradiation) and 50% and 75% (partial-body irradiation) under densely ionising radiation. For reference, each of these exposure scenarios are labelled as "**D1**", "**D2**" and "**D3**" respectively in Tables 1.1, 1.2 and 1.3. In each case, frequencies of dicentrics and centric rings are recorded. Data are also available from [76] and correspond to a culture time of 72h.

Although the maximum number of dicentrics observed in a single cell is 7, larger counts are possible but it is fairly uncommon to score more than 10 dicentrics per cell. By contrast, the range of foci counts will typically be on a wider scale. We will make use of three H2AX datasets. The first dataset entitled "**PHE-Foci1**" is part of a much larger dataset whereby focus counts are based on small examined samples of 200 cells for $4h$ and $24h$ timepoints. Blood was divided into 30/70%, 40/60%, 60/40%, 80/20% and 100/0% irradiated/non-irradiated ratios. For the second dataset "**PHE-Foci2**", foci were scored over 1000 cells, counted manually $4h$ post exposure. The proportions of blood irradiated were 30%, 60%, and 100%. In both datasets, foci were scored in an

| (**D1**) | Dicentrics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Dose | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Total | Sample Size |
| 0.1 | 2281 | 130 | 21 | 1 | 0 | 0 | 0 | 0 | 175 | 2433 |
| 0.3 | 847 | 127 | 19 | 6 | 1 | 0 | 0 | 0 | 187 | 1000 |
| 0.5 | 567 | 165 | 49 | 16 | 2 | 0 | 0 | 0 | 319 | 799 |
| 0.7 | 356 | 167 | 62 | 9 | 5 | 1 | 0 | 0 | 343 | 600 |
| 1 | 169 | 131 | 72 | 18 | 9 | 0 | 0 | 1 | 372 | 400 |

Table 1.1 Dicentric distribution under whole-body irradiation.

| (**D2**) | Dicentrics | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dose | 0 | 1 | 2 | 3 | 4 | 5 | Total | Sample Size |
| 0.5 | 875 | 88 | 30 | 7 | 0 | 0 | 169 | 1000 |
| 0.7 | 679 | 88 | 23 | 8 | 1 | 1 | 167 | 800 |
| 1 | 480 | 75 | 27 | 13 | 5 | 0 | 188 | 600 |

Table 1.2 Dicentric measurements under 50% partial-body exposure conditions.

| (**D3**) | Dicentrics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Dose | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Total | Sample Size |
| 0.5 | 633 | 118 | 37 | 10 | 1 | 1 | 0 | 231 | 800 |
| 0.7 | 455 | 98 | 37 | 9 | 1 | 0 | 0 | 203 | 600 |
| 1 | 263 | 88 | 36 | 11 | 1 | 0 | 1 | 203 | 400 |

Table 1.3 Dicentric measurements under 75% partial-body exposure conditions.

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | 0.75 Gy | | | 1.5 Gy | | | 3 Gy | | |
| 2 | Control | 30% | 60% | 100% | 30% | 60% | 100% | 30% | 60% | 100% |
| 3 | 0 | 4 | 1 | 3 | 9 | 4 | 7 | 0 | 0 | 10 |
| 4 | 0 | 0 | 1 | 3 | 0 | 1 | 7 | 0 | 0 | 4 |
| 5 | 0 | 0 | 1 | 2 | 0 | 5 | 3 | 0 | 0 | 5 |
| 6 | 0 | 0 | 1 | 3 | 0 | 0 | 1 | 0 | 0 | 10 |
| 7 | 0 | 5 | 0 | 5 | 1 | 0 | 5 | 1 | 0 | 11 |
| 8 | 0 | 3 | 4 | 2 | 6 | 0 | 5 | 12 | 0 | 4 |
| 9 | 0 | 0 | 3 | 3 | 1 | 7 | 7 | 0 | 0 | 9 |
| 10 | 0 | 0 | 1 | 1 | 0 | 4 | 7 | 0 | 0 | 12 |
| 11 | 0 | 0 | 3 | 7 | 0 | 0 | 7 | 0 | 0 | 13 |
| 12 | 0 | 2 | 4 | 5 | 0 | 0 | 8 | 9 | 0 | 9 |
| 13 | 0 | 0 | 6 | 4 | 6 | 0 | 7 | 8 | 0 | 11 |
| 14 | 0 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 7 | 12 |
| 15 | 0 | 0 | 2 | 2 | 0 | 6 | 6 | 0 | 10 | 9 |
| 16 | 0 | 0 | 0 | 1 | 0 | 0 | 5 | 0 | 13 | 15 |
| 17 | 0 | 3 | 0 | 3 | 0 | 0 | 4 | 13 | 14 | 12 |
| 18 | 0 | 0 | 2 | 5 | 0 | 0 | 7 | 0 | 12 | 11 |
| 19 | 0 | 5 | 1 | 3 | 0 | 9 | 4 | 9 | 15 | 13 |
| 20 | 0 | 0 | 2 | 4 | 0 | 4 | 3 | 0 | 14 | 11 |
| 21 | 0 | 0 | 0 | 2 | 1 | 6 | 3 | 0 | 11 | 11 |
| 22 | 0 | 5 | 0 | 6 | 2 | 8 | 3 | 0 | 17 | 7 |
| 23 | 0 | 0 | 1 | 3 | 0 | 0 | 9 | 0 | 10 | 6 |
| 24 | 0 | 0 | 5 | 4 | 9 | 6 | 9 | 12 | 2 | 15 |
| 25 | 0 | 1 | 3 | 5 | 1 | 8 | 7 | 8 | 0 | 13 |
| 26 | 0 | 0 | 1 | 6 | 0 | 8 | 4 | 0 | 0 | 7 |
| 27 | 0 | 0 | 0 | 5 | 1 | 9 | 8 | 0 | 0 | 12 |
| 28 | 2 | 0 | 0 | 2 | 0 | 1 | 9 | 11 | 15 | 10 |

Fig. 1.3 A screenshot showcasing a segment of the raw data included in the PHE-Foci2 dataset.

in-vitro setting at Public Health England following irradiation of blood lymphocytes with 250kVp X-rays using dose levels of 0, 0.75, 1.5 and 3Gy. For clarity, the results from the 0Gy dose make up the control data i.e. foci scored under 0% exposure/no irradiation or background foci.

For the reasons stated above, it is not particularly convenient for raw H2AX data to be provided in the same manner as dicentric data (as in Tables 1.1-1.3 or similar) but rather in electronic format. To illustrate, an example of inputting PHE-Foci2 raw data using Excel software is shown in Fig 1.3. Histograms to show the breakdown of foci measurements in each sample at 4h post-exposure are displayed in Figures 1.4 and 1.5. Immediately we identify that the PHE-Foci2 irradiated samples provide larger foci counts which can only be concretely explained through a change of technology in the scoring process. If both datasets were conducted under the exact same conditions then it is reasonable to suggest that some counts from the PHE-Foci1 dataset should be in proximity to 20 (ignoring for now the anomalous cell of 37 scored foci).

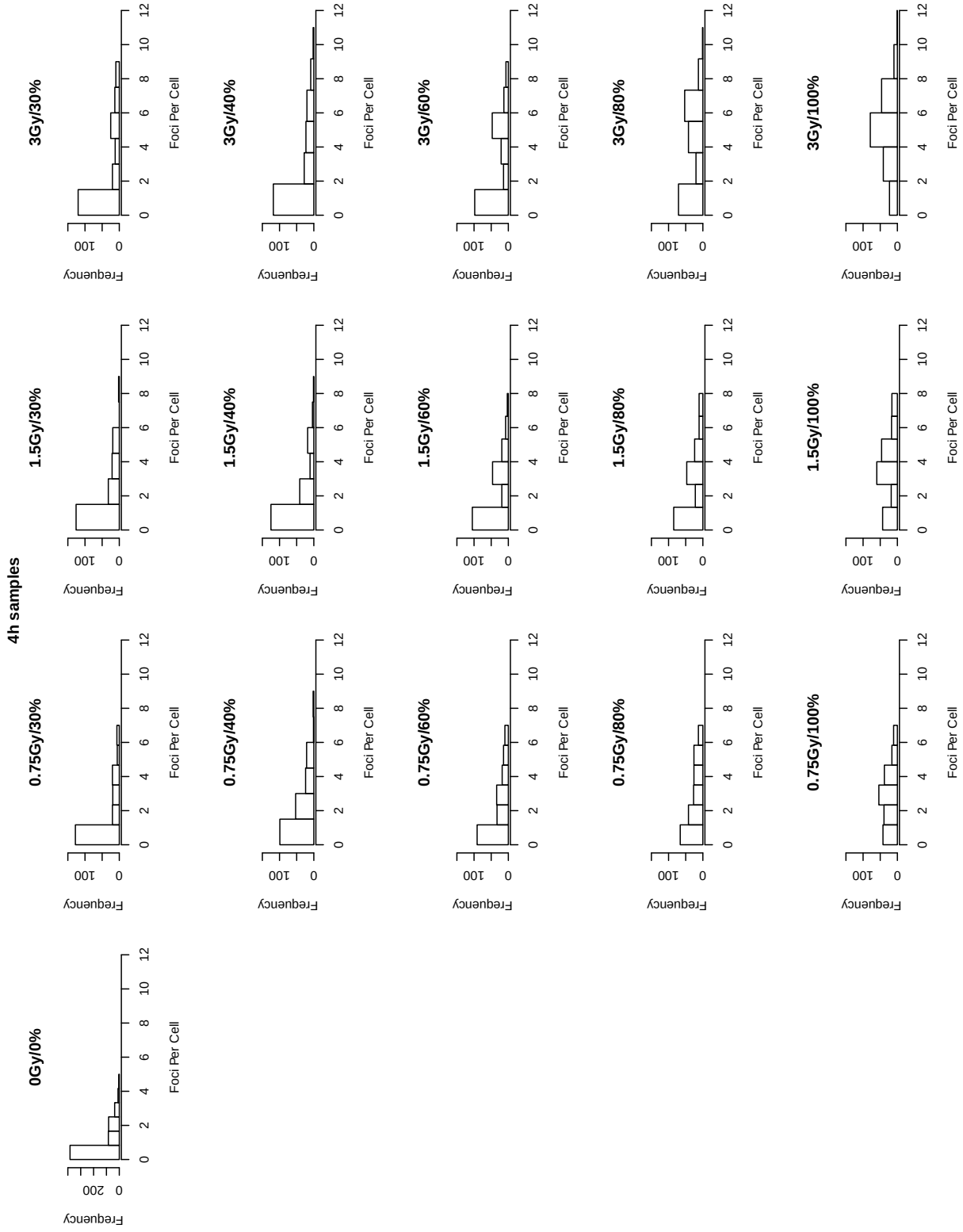



Fig. 1.4 Foci distributions of 4h PHE-Foci1 samples.

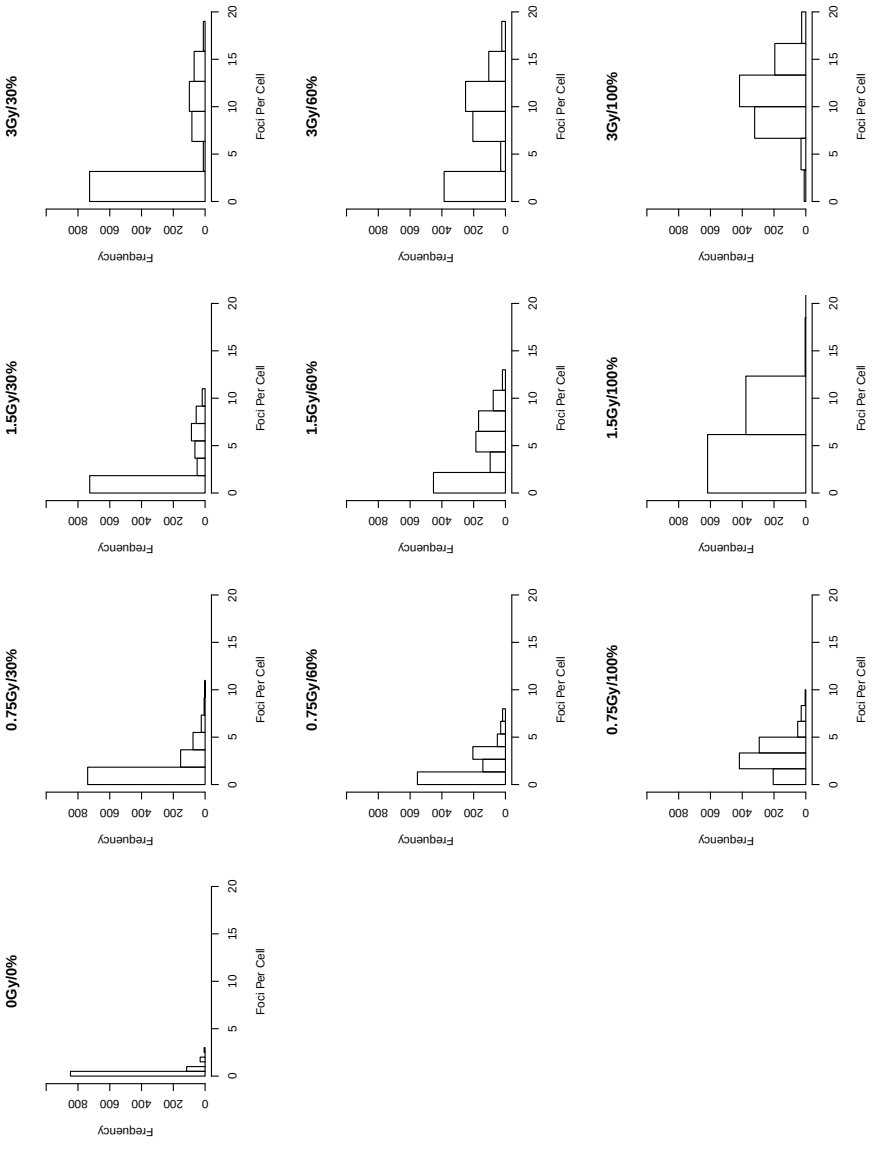



Fig. 1.5 Foci distributions of 4h PHE-Foci2 samples (note: the 1.5Gy/100% sample contains a single cell consisting of 37 foci but x-axis range is chosen to reflect the second highest recorded frequency of 20 foci).

Taking the maximum recorded count in each PHE-Foci1 sample, we investigated the proportion (out of 1000) of cells in the PHE-Foci2 samples (of the same exposure fraction) which were greater than this count. For the 0.75Gy samples, this translates to 0.7% (30%), 0.4% (60%) and 1.2% (100%). Since these proportions are very small, then in a sample of 200 we would expect only 1 or 2 count(s) to be greater than 7. Aside from the 1.5Gy/30% sample, the proportions for the remaining samples range from 9.6% to 37.9%. Given the same number of cells analysed for both datasets, it could be speculated that their distributions become alike, however one can only proceed with what is available to them.

The aggregation of count data prior to analysis or modelling is a very common procedure in several fields, including for instance the aggregation of clickstream data in e-commerce [42], or of species counts in ecology [41]. Furthermore, in biodosimetry, it is common to aggregate counts of certain biomarkers over samples of blood cells and use the aggregated count for the estimation of dose-response curves, or the estimation of dose given an existing curve.

In order to establish dose-response calibration curves, laboratory experiments are carried out where blood samples are exposed to known degrees of radiation. The data arising from a series of such experiments conducted at the Bundesamt für Strahlenschutz (BfS), Germany, are displayed in Figures 1.6 and 1.7. Henceforth we will refer to this dataset as "**BfS-Foci**". For the production of this data, whole blood samples were irradiated with one of six design doses (0.1, 0.2, 0.3, 0.4, 0.5 and 1Gy), always with 195kV X-radiation. One hour after exposure, blood samples consisting of approximately 2000 cells were then placed on slides under an immunoflourescence microscope, and the number of foci on each slide was counted in a semi-automatic way using MetaCyte software. (Additional information on the generation of this dataset is deferred to Appendix A.4).

In total, measurements from 116 slides are available, corresponding to a total of 233220 frequencies of foci per cell. Fig. 1.6 gives an excerpt of the raw data, in the form of a frequency distribution of foci counts for three specific slides. One sees clearly how the distribution of the foci counts is shifted to the right for increasing doses, in a similar manner to Figures 1.4 and 1.5, underlining their suitability as a radiation biomarker (note again that all cells on a given slide always share the same design dose). The full BfS-Foci dataset is displayed in Fig. 1.7 in aggregated form, with each point corresponding to the mean foci count for a specific slide. From this one can deduce some sort of empirical dose-response relationship, which appears roughly linear over a considerable dose range, noting a saturation effect [74] for higher doses.

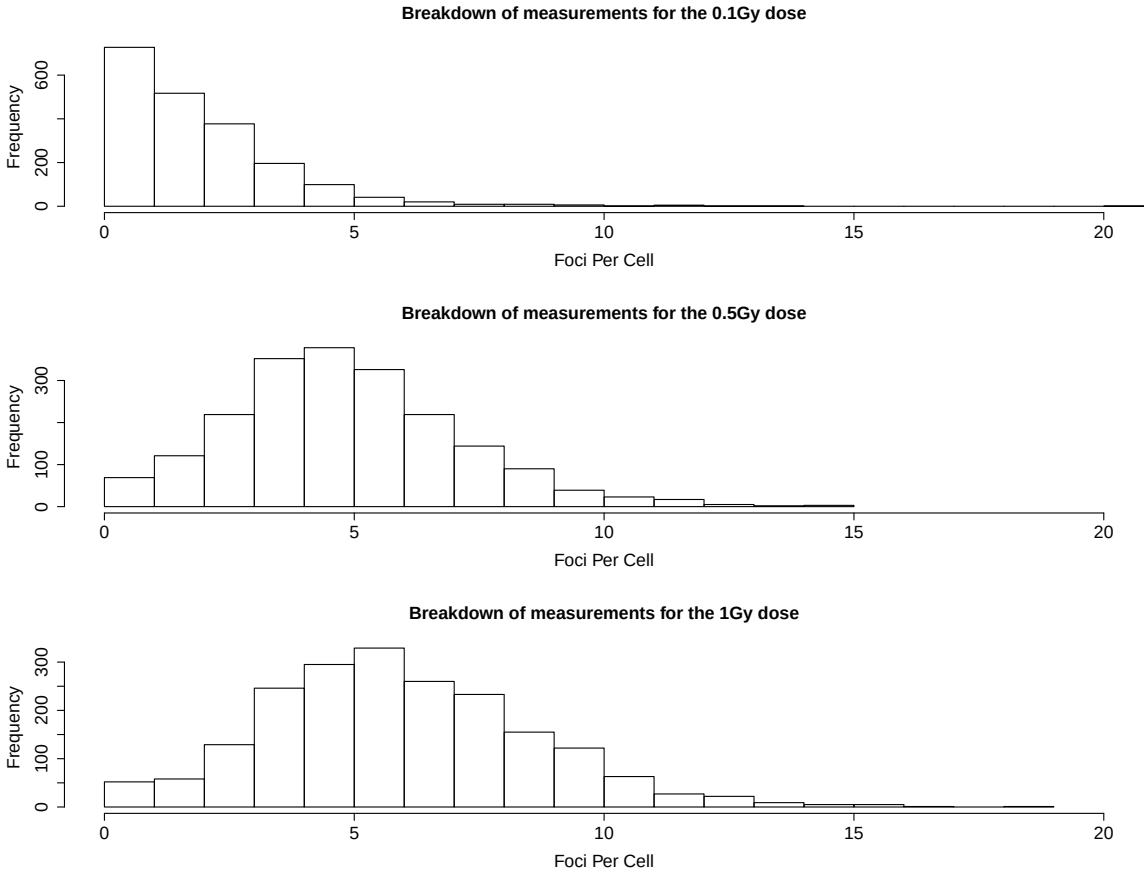



Fig. 1.6 Distribution of the number of observed foci, for three selected slides from BfS-Foci dataset with dose levels 0.1Gy, 0.5Gy and 1Gy, respectively.

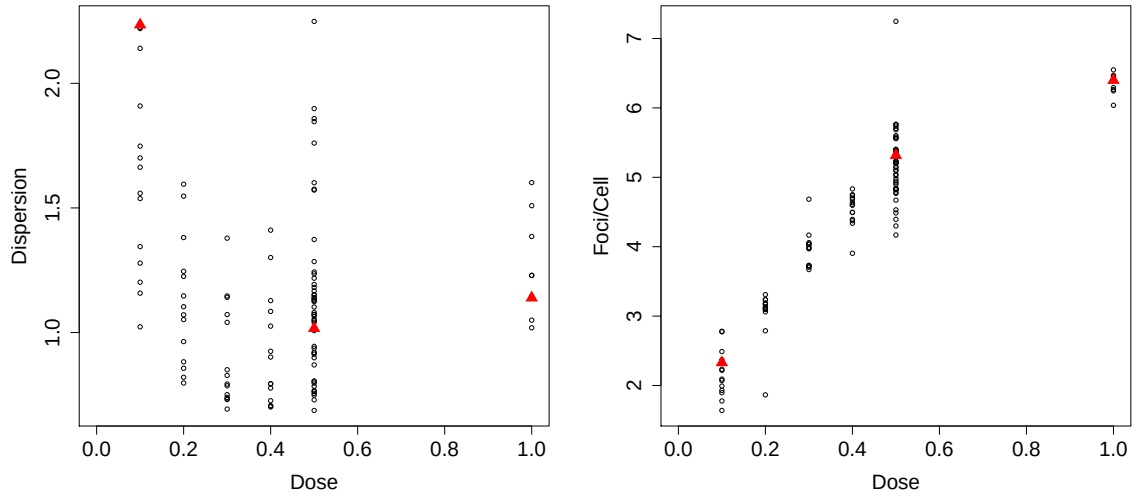



Fig. 1.7 BfS-Foci slide-wise dispersions (left) and foci yields (row-means) (right) recorded for various levels of dose. The three points highlighted as triangles indicate the specific slides which have been displayed in Fig. 1.6.

## 1.3 Outline of the thesis

In this chapter we have discussed the advantages and disadvantages of both the H2AX and dicentric assay, as well-established biomarkers in the field of biodosimetry. The main focus of this thesis will be the H2AX assay, however some of our analysis will make use of the PHE-Dicentric data presented in Section 1.2. In the following Chapter we will review some of the existing approaches to dose estimation, based on frequentist and Bayesian methods. Some work on the implementation of the zero-inflated Poisson and NB1 models using MCMC (via the R package `rjags`) can be found in [30], where we concluded that there is no "strong" preference in selecting between a frequentist or Bayesian framework for the purpose of fraction estimation. We note that these results were conditional to the use of a uniform prior which suggests any exposure scenario is equally likely. Although beyond the scope of our work, the choice of prior could be updated, with weights or probabilities assigned to certain exposures, to reflect the case that some partial-exposures are more likely than others. We note that this thesis will draw focus from the frequentist perspective.

In Chapter 3, we will introduce separably the concepts of dispersion and zero-inflation which will allow us to critically assess the adequacy of the Poisson in describing a scored foci or dicentric distribution. This will serve as a preliminary to Chapter 4, in which we define statistical count data models for handling overdispersion in a formal manner as well as the notion of likelihood and maximum likelihood. We conclude this chapter through the analysis of calibration data, exploring in detail the behaviour of dispersion and zero-inflation against dose level and fraction of exposure.

A distinct property of maximum likelihood is that it allows the standard likelihood tests to be implemented. In Chapter 5, we discuss further how the overdispersion parameter (estimated in Chapter 4) can be used as information against the Poisson model via the likelihood ratio and Wald test. We then proceed to compare these tests with score tests proposed by Dean and Lawless which have the advantage over the likelihood ratio and Wald tests in that they only require the parameter estimated under the null hypothesis. In addition, model selection based on information criterion will be employed.

In the event that a calibration curve has been constructed based on a range of dose values involving multiple experiments or slides, it is convenient for a laboratory to supply such data in aggregated format. In Chapter 6, we make full use of the BfS-Foci dataset which conveniently allows us to compare the dispersions in the raw and aggregated data. We will show through both theory and simulation that relatively small deviations from the independence assumption in the raw data (say, the presence of strings of correlated observations) can increase the dispersion of the aggregated data dramatically. Although this is not entirely novel, the behaviour of dispersion estimates under aggregation is under-reported, almost certainly in the field of biodosimetry. We finish this chapter by highlighting the pros and cons of aggregated data with respect to raw data.

Following a potential radiation incident, a clinician will often be provided with only one exposed patient's blood sample. As opposed to using calibration data (as in Chapters 4 and 5), Chapters 7 and 8 are focused on dose and fraction estimation in the case of a single sample. We begin by outlining the contaminated Poisson method before making a novel attempt to update this method in order to account for additional overdispersion which is not attributable to zero-inflation. Part of this work is discussed in [31]. We will see that certain data characteristics, for example an excessive frequency of low (non-zero) counts, can create problems for these methods. The aim of Chapter 8 is then to define alternative procedures or steps which can be taken in such circumstances to improve estimates.

Finally in Chapter 9, we conclude this thesis. We discuss our findings and issues while investigating the problem.

# Chapter 2

# Current procedures for dose estimation

One of the biggest challenges in biological dosimetry is the satisfactory conversion of a measured quantity of radiation damage, such as a dicentric or foci yield, into an estimate of dose. All biodosimetric methods require a calibration curve in order to translate the observed yield of damage in cells into a radiation dose estimate. In addition, the choice of radiation biomarker used to quantify the contracted dose through the caused cellular damage, is important in assessing the radiation sensitivity in a patient blood sample. The aim of this chapter is to review and compare frequentist and Bayesian processes of arriving at a whole-body dose estimate and, given this estimate, how exactly its uncertainty can be quantified.

## 2.1 Generalised linear models

We begin this chapter by outlining some terminology. We refer to a set of aberration counts (constituting a specific histogram such as in Figures 1.4, 1.5 and 1.6) as a *slide*. The $j$-th observation in the $i$-th slide is denoted as $y_{ij}$, for $k$ slides with respective size $n_i$, $i = 1, \dots k$. Averaging over the $i$-th slide, we obtain the means $y_i = \sum y_{ij}/n_i$, which are also referred to as *yields* in the dosimetry literature. The convention to speak of *slides* and *yields* is simply with reference to the data considered, and is not implying a restriction of the validity of the results to this particular field of application.

Fixing terms, we can relate the yields to dose via the expression:

$$y_i = E(y_{ij}|x_i) = \beta_0 + \beta_1 x_i + [\beta_2 x_i^2] \tag{2.1.1}$$

where, under this framework, the predictor $x_i$ (here, representing the design doses) will never depend on $j$. For whole-body exposure, the distribution of foci (and dicentrics) among the irradiated cell population is assumed to be Poisson. The modelling strategy for the Poisson distribution is set out more formally in the following section.

For many years, the usage of generalised linear models (GLMs) [71] has become standard for the construction of dose-response calibration curves. By definition, GLMs are a class of models that generalise linear regression where the response variable is expected to follow an exponential dispersion family (EDF) with mean $\mu$. In general, there are 3 components needed to define a GLM:

1. Random component:

   Given covariates $x_i$, the responses $y_{ij}$ are iid with an EDF density of the form

   $$f(y;\theta,\phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.1.2}$$

   ,

   where $\theta \in \mathbb{R}$ is the 'natural parameter', $\phi > 0$ is the 'dispersion parameter' (discussed in further detail in the following chapter) and $a, b : \mathbb{R} \to \mathbb{R}$ and $c : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ are functions.

2. Systematic component:

   $\eta_i = x_i'\boldsymbol{\beta}$, the linear predictor, where $\boldsymbol{\beta}$ is the vector of unknown regression parameters.

3. Parametric link component:

   The link function $g(\mu_i) = \eta_i = x_i'\boldsymbol{\beta}$ combines the linear predictor with the mean $\mu_i$ of $y_{ij}$. Here, the canonical link function is used, so that $\boldsymbol{\theta} = \boldsymbol{\eta}$ holds.

In our context, classical linear regression is deemed inappropriate for fitting dose–response curves since responses are the mean of Poisson counts and the assumption of homoscedasticity is violated due to the biological process which leads to dependence of the variance on the dose. As a first model for count data, we initially consider the Poisson regression.

## 2.2 Poisson models

While models based on the Gaussian distribution, such as the common linear model, are usually not applicable to count data, this is not a big obstacle as count data regression models are now well developed. The most basic of all count data models is the Poisson model, which postulates

$$y_{ij} \sim \text{Pois}(\mu_i), \tag{2.2.1}$$

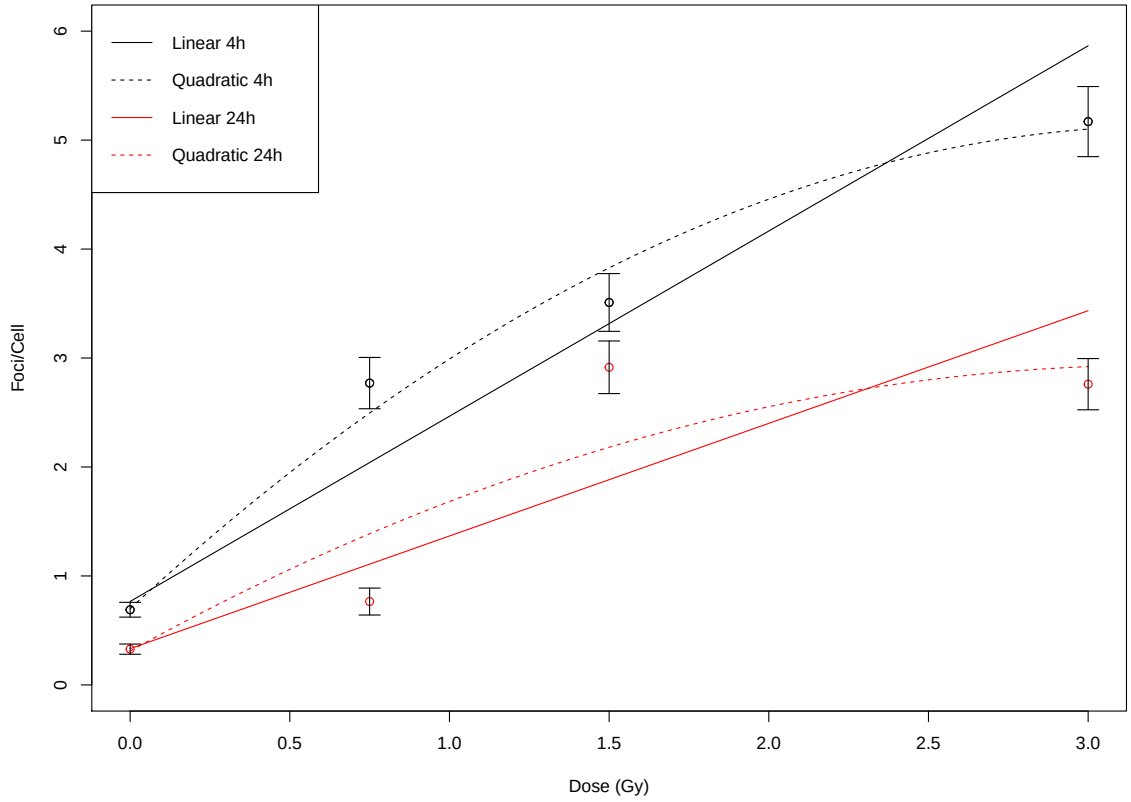



Fig. 2.1 Poisson linear (solid) and quadratic (dashed) calibration curves fitted to 4h (black) and 24h (red) PHE-Foci1 data. Error bars represent ±2× Poisson sampling error [see Appendix A.2].

that is

$$f(y_{ij}|x_i) = e^{-\mu_i}\frac{\mu_i^{y_{ij}}}{y_{ij}!},$$

where

$$\mu_i = g^{-1}(x_i'\boldsymbol{\beta}), \tag{2.2.2}$$

with $\mu_i > 0$, for some link function $g$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, [\beta_2])' \in \mathbb{R}$ as in (2.1.1). This means we assume that all observations from a particular sample or slide will share the same predictors, and hence the same $\mu_i$.

The Poisson density can be rewritten in the form of the exponential family (2.1.2) as $f(y_i) = \exp\{-\mu + y\ln(\mu) - \ln(y!)\}$ with $a(\phi) = 1, \theta = \ln(\mu), b(\theta) = \mu$ and $c(y, \phi) = -\ln(y!)$. The canonical link function typically used is the log-link function $\eta = \ln(\mu)$, however, in the context of dosimetry, one will usually use the identity link function $g(\mu) = \mu$. This choice is motivated by physical considerations and the shape of the dose-response curve, despite the fact that the log-link is, from a statistical viewpoint, a natural choice for count data. For the sole purpose of fitting dose-response curves

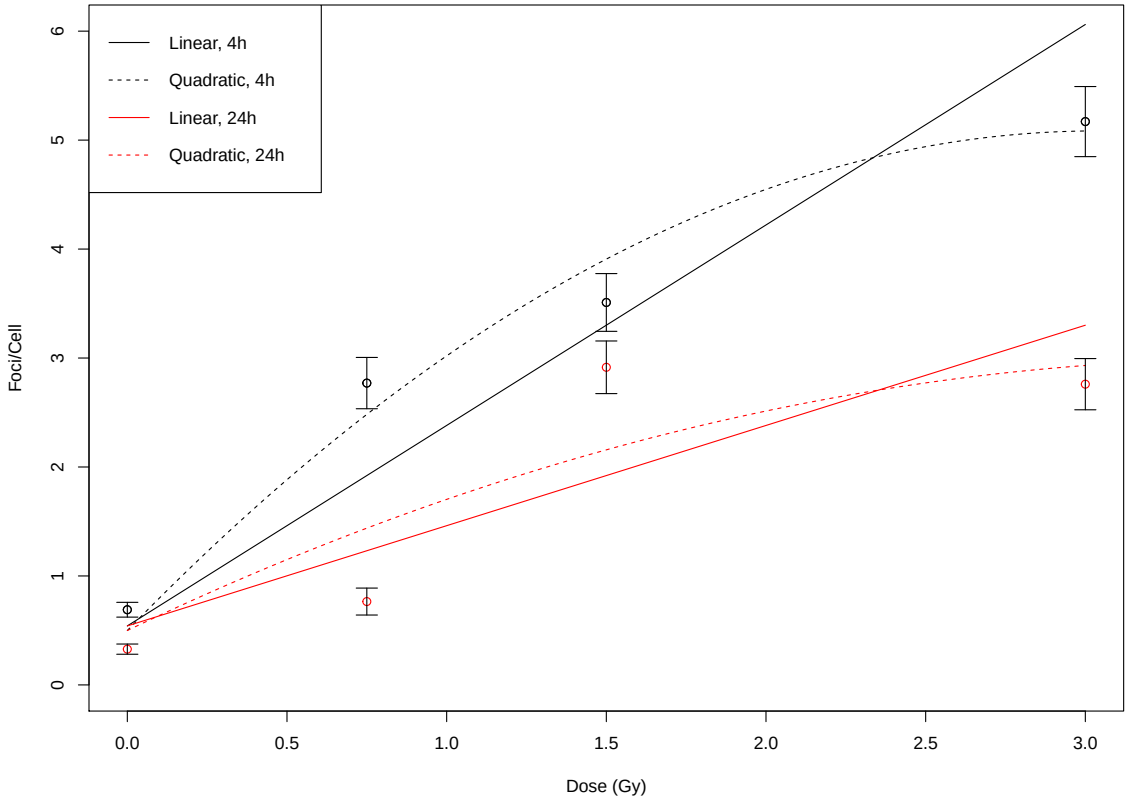



Fig. 2.2 Linear (solid) and quadratic (dashed) combined calibration curves.

to calibration data (typically whole-body), it is standard to relate the mean yield of aberrations, $\lambda_i$, to dose $x_i$ via the linear [or quadratic] model [28]

$$\lambda_i = E(y_{ij}|x_i) = \beta_0 + \beta_1 x_i + [\beta_2 x_i^2]. \tag{2.2.3}$$

For clarity, when we speak of calibration data, we refer to data arising from multiple experiments conducted using various doses as shown in Table 1.1 for example. In some cases, this may also include the 0Gy control samples i.e. aberrations scored under no exposure, as depicted in both Figures 2.1 and 2.2 for the PHE-Foci1 4h and 24h whole-body calibration data.

It is clear from Figures 2.1 and 2.2 that counts of $\gamma$-H2AX increase with level of dose but decrease with time. The presence of foci indicates that DSBs have occurred and are initiating a cellular response to repair the damage. During the early stages of DSB repair, the number of H2AX foci per cell increases as the DNA damage response is activated and the foci accumulate at sites of DNA damage. It is believed that measurements of $\gamma$-H2AX peak at approximately 30 minutes following exposure. Subsequently, as repair mechanisms begin to repair the DSBs, the foci yield gradually decreases (hence the reason we see smaller yields at 24h post exposure) before returning

to baseline levels within 24-48 hours. Due to the variability in foci counts over time, calibration curves are commonly constructed separately for various times post exposure, as in Fig 2.1.

Assuming the data $y_{ij}$ to be conditionally independent given $x_i$, the model likelihood can be written as

$$L = \prod_{i,j} f(y_{ij}|x_i) = \prod_{i,j} e^{-\mu_i} \frac{\mu_i^{y_{ij}}}{y_{ij}!} \propto \prod_i e^{-n_i\mu_i} \mu_i^{\sum_j y_{ij}}.$$

We recall that for inferential purposes concerning the model parameters, the required information for the likelihood is fully provided by the sums $s_i = \sum_j y_{ij}$, or equivalently by the yields $y_i = s_i/n_i$. This property, known as 'sufficiency', implies that the aggregated data (i.e. when referring either to $y_i$ or $s_i$) contain sufficient information for inference on $\mu_i$, and, hence, $\boldsymbol{\beta}$. Notably, this does not only hold for the parameter estimates but also their standard errors; in other words, given the aggregated data, no improvement in either accuracy or precision is possible by considering the raw data.

One can interpret the intercept parameter in (2.2.3) as the 'background yield' constant, that is, the expected yield under zero dose. However, in the absence of radiation exposure it is meaningless to speak of 'time after irradiation', hence this constant should be identical for each calibration curve. Assuming data are obtained under the same experiment conditions, we can define a combined model of the form:

$$E(y_{ij}|x_i) = \beta_0 + \beta_1 x_i + [\beta_2 x_i^2] + \beta_3 x_i \mathbf{1}\{time = 24h\} + [\beta_4 x_i^2 \mathbf{1}\{time = 24h\}], \quad (2.2.4)$$

displayed in Figure 2.2. As a further extension, if the degree of exposure is known then it is possible to extend (2.2.4) such that

$$E(y_{ij}|x_i) = \beta_0 + \rho(\beta_1 x_i + [\beta_2 x_i^2] + \beta_3 x_i \mathbf{1}\{time = 24h\} + [\beta_4 x_i^2 \mathbf{1}\{time = 24h\}]), \quad (2.2.5)$$

where $\rho \in [0, 1]$ with $\rho = 0$ and $\rho = 1$ representing non-irradiation and whole-body exposure respectively.

### 2.2.1 Poisson mean-variance relationship

A main principle of the Poisson regression is that of equidispersion. For all $i$ and $j$,

$$\mathrm{Var}(y_{ij}|x_i) = E(y_{ij}|x_i) = \mu_i, \quad (2.2.6)$$

implying trivially that

$$\frac{\mathrm{Var}(y_{ij}|x_i)}{E(y_{ij}|x_i)} = 1. \tag{2.2.7}$$

Another important characteristic of the Poisson model is that the equidispersion carries over from the raw to the aggregated data model,

$$E(s_i|x_i) = n_i g^{-1}(x_i'\boldsymbol{\beta}). \tag{2.2.8}$$

This property, along with the sufficiency property, makes a compelling case for the use of the aggregated data in Poisson models: They contain all required information but require less storage space, less computational time to fit the models, and allow for simplified data display.

In application to practical radiation biomarker data, the equidispersion property will most often be violated. In such cases, a Poisson model may underestimate parameter uncertainty and therefore overstate the significance of those parameters. For the avoidance of misleading inference, we require models, for example as outlined later in Chapter 4, which do not possess the same restriction of mean-variance equality.

## 2.3 Frequentist methodology

We consider now the scenario that a given calibration curve (2.1.1) is available. A blood sample has been taken from a potentially exposed individual, and a number $n^*$ of cells of this sample have been examined. These $n^*$ cells deliver a total focus count $s^*$ and hence a yield $y^* = s^*/n^*$. From (2.1.1), the whole-body equivalent dose estimate can be obtained through inverse regression i.e.

$$x^* = \frac{y^* - \hat{\beta}_0}{\hat{\beta}_1} \tag{2.3.1}$$

or for the quadratic case (as commonly used with cytogenetic biomarkers):

$$x^* = \frac{-\hat{\beta}_1 \pm \sqrt{\hat{\beta}_1{}^2 - 4\hat{\beta}_0\hat{\beta}_2}}{2\hat{\beta}_2}. \tag{2.3.2}$$

For the $\gamma$-H2AX assay, it is not convenient to include the quadratic term (for reasons discussed later in Chapter 4) hence the following section is based on (2.3.1).

## 2.4 Uncertainty quantification

Only recently has the question of how to assess quantitatively the uncertainty related with a radiation dose estimate started gaining interest. When we speak of uncertainty, we refer to the variability resulting from both the sampling process and fitted calibration curve. The aim is to ultimately express uncertainty in terms of a confidence interval and it is standard practice to calculate 95% limits. The 95% confidence limits define an interval that will encompass the true dose on at least 95% of occasions. The following methods can be used to calculate uncertainty for a given dose estimate:

1. **Delta approximation** (*'Multibiodose method'*)

   Firstly, the uncertainty (expressed in the form of standard errors) attached to a dose estimate $x^*$ can be decomposed via [see Appendix A.5]

$$SE^2(x^*) = \left(\frac{\partial x^*}{\partial \hat{\beta}_0}\right)^2 SE^2(\hat{\beta}_0) + \left(\frac{\partial x^*}{\partial \hat{\beta}_1}\right)^2 SE^2(\hat{\beta}_1) + \left(\frac{\partial x^*}{\partial y^*}\right)^2 SE^2(y^*). \quad (2.4.1)$$

   The partial derivatives in (2.4.1) can be worked out to be

$$\frac{\partial x^*}{\partial \hat{\beta}_0} = -\frac{1}{\hat{\beta}_1}$$

$$\frac{\partial x^*}{\partial \hat{\beta}_1} = \frac{\hat{\beta}_0 - y^*}{\hat{\beta}_1{}^2}$$

$$\frac{\partial x^*}{\partial y^*} = \frac{1}{\hat{\beta}_1}.$$

   .

   The quantities in (2.4.1) are immediately known from either the calibration curve or sample information except the error associated with the yield, $SE(y^*)$. For $y_{ij}$ assumed to be Poisson distributed, this quantity is equivalent to the Poisson sampling error [Appendix A.2].

2. **Merkle's method**

   Merkle's method [73] assumes the error of both the aberration yield and calibration curve coefficients is of Poisson nature. The method is detailed in IAEA manual and is well-established for use with the dicentric biomarker. In summary, it involves calculating upper and lower 95% confidence limits on the yield ($y_U^*$ and $y_L^*$) using the following equations:

$$y_U^* = \frac{0.5\chi^2_{0.025,2y^*n^*+2}}{n^*} \qquad y_L^* = \frac{0.5\chi^2_{0.975,2y^*n^*}}{n^*}. \tag{2.4.2}$$

The upper and lower confidence limits on the curve are then calculated according to

$$y^*_{U/L}(x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^* \pm CF\sqrt{SE^2(\hat{\beta}_0) + SE^2(\hat{\beta}_1)x^{*2}}, \tag{2.4.3}$$

where $CF = 2.45$ is the regression 95% confidence factor of the Chi-squared distribution with 2 degrees of freedom (linear model). To obtain a 95% confidence interval of the "true" dose, we determine the dose at which the yield $y_L^*$ intersects with $y_U^*(x^*)$ to provide the lower confidence limit on the dose, and evaluate the dose at which $y_U^*$ crosses with $y_L^*(x^*)$ to give the upper confidence limit on the dose.

A proposed refinement to this method to reduce possible overestimation of uncertainty is to instead use an 83% confidence interval. However, there is little point in shrinking the interval size without first checking whether this method works for the $\gamma$-H2AX histone. In order for the method to be successful, the expectation was that almost 95% of the intervals produced should contain the real doses. An attempt to apply Merkle's method for generating 95% confidence intervals to $\gamma$-H2AX data with known real doses has previously been made in [36]. They concluded that the method was insufficient, with a very low success rate of only 7% of measured yields providing intervals containing the real dose.

3. **IAEA simplification**

   This method also relies on using the upper and lower confidence limits on the yield, as computed in (2.4.2). However, in comparison to Merkle's method, the dose confidence limits are found by simply substituting $y_U^*$ and $y_L^*$ directly into the dose estimator (2.3.1).

4. **Monte Carlo propogation of errors**

   A distinct advantage of Monte Carlo (MC) simulation is that uncertainty calculations can be accomplished by non-statisticians using standard spreadsheet applications (such as Excel) rather than requiring technically demanding mathematical procedures. To apply the method, we assume the input quantities $(y^*, \hat{\beta}_0, \hat{\beta}_1)$ are normally distributed. Based on these distributions, trial values can be generated for each of these input variables using a combination of the NORMINV (for calculating Gaussian-distributed values) and RAND (for generating pseudo-random values) functions, according to the procedure outlined in [33]. We note in the instance a negative value is drawn for $\hat{\beta}_0$, a corrected value of $\hat{\beta}_0 = 0.001$ is used. The estimated dose $x^*$ is then calculated using (2.3.1). The

above process constitutes a single MC simulation. The idea is to perform multiple simulations, providing an overall mean and standard deviation to construct a 95% confidence interval.

## 2.5 A probabilistic approach

In parallel to the classical, frequentist convention, Bayesian methods are becoming increasingly popular in the field of biological dosimetry [46, 17, 4]. Key to the Bayesian concept is the application of the inversion theorem in its continuous version, i.e.

$$P(x^*|y^*) = \frac{P(y^*|x^*)P(x^*)}{\int P(y^*|x^*)P(x^*)dx^*}$$

Thus, the posterior dose distribution (or calibrative dose density), $P(x^*|y^*)$, scales with the product of the likelihood (or predictive density) and the prior $P(x^*)$:

$$P(x^*|y^*) \propto P(y^*|x^*)P(x^*).$$

With respect to uncertainty analysis, the Bayesian approach does not require additional considerations, since the resulting distribution $P(x^*|y^*)$ inherently provides quantification of the uncertainty within the dosimetric model. Consequently, Bayesian uncertainty intervals for the dose parameter are accurate.

Apart from the intrinsic inclusion of uncertainty within the posterior model, additional information besides the number of aberrations can be used through the chosen prior distribution(s). The choice of the prior could be sensitive, since well chosen, informative priors should guide noisy data towards the true dose, whereas incorrect priors may drive the estimate away from the true dose. Higueras also showed that if an appropriate prior is applied, the actual choice of prior in fact does not greatly impact the overall dose assessment in some scenarios [46].

Higueras et al.[45] also discussed the reasonable set-up of Poisson and compound Poisson models (Neyman A, negative binomial, Hermite) with parameterisation based on using the "dispersion index" (which we will discuss in the following Chapter) for biodosimetry. Furthermore, the authors also presented a guide for analysis of partial-body exposure for a zero-inflated Poisson model [46].

For biological dosimetry, it can be concluded that the Bayesian methodology is certainly coherent, but at the same time it is far more technically challenging than the frequentist dose and uncertainty assessment methods currently recommended and used by most practitioners [53]. In particular, the potential pitfall of incorrectly chosen priors needs careful consideration. Software-based solutions, however, can certainly help bridge the gap between the necessary mathematical skills and the users.

### 2.5.1 Software and web-based applications

Over the years, several software packages have been developed to help assist with curve fitting and dose estimation. Until recently, the two most frequently used software packages were CABAS [25] and Dose Estimate [2]. Due to their simplicity, there is an increasing popularity in the usage of web-based applications, for example DoseEstimateH2AX [28] and BiodoseTool [43]. In addition to curve fitting and deriving dose estimates together with the associated uncertainty and probability parameters, there are also statistical tools which allow for partial body calculations, genome equivalents etc.

Since they allow inclusion of previous information about the circumstances of exposure [5], more user-friendly software platforms based on Bayesian modeling have been developed to estimate radiation exposures including CytobayesJ and radir [45]. The latter package is able to plot calibrative dose densities as well as provide relevant summary statistics including best dose estimate, standard deviation and credibility interval. Figure 2.3 provides the resulting calibrative dose density upon application to dataset D1, omitting the 0.5Gy row as test data. This can be reproduced in RStudio as follows

```
dose <- c(0.1,0.3,0.5,0.7,1)
freq <- matrix(c(2281,847,567,356,169,130,127,165,167,131,21,19,49,62,
72,1,6,16,9,18,0,1,2,5,9,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1),5,8)
d <- dim(freq)[1]
ndic <- dim(freq)[2]-1
ncel <- rowSums(freq)
X <- as.vector(freq%*%(0:ndic))
dic <- dosevec <- numeric()
for(i in 1:d){
  for (j in 0:ndic){
    dic <- c(dic,rep(j,freq[i,j+1]))
    dosevec <- c(dosevec,rep(dose[i],freq[i,j+1]))
  }}
datavecD1 <- data.frame(dic,dosevec,dosevec2=dosevec^2) # raw data
dataggrD1 <- data.frame(X,ncel,dose,dose2=dose^2) # aggregated data
fit <- glm(X~-1+I(ncel)+ I(ncel*dose) + I(ncel*dose^2), family=
poisson(link = "identity"), data=dataggrD1[-3,]) # Poisson curve with
0.5Gy row removed

install.packages("radir")
library(radir)
f <- expression(b0+b1*x+b2*x^2)
pars <- c("b0","b1","b2")
beta <- fit$coef
cov <- summary(fitb2)$cov.unscaled
ex1 <- dose.distr(f, pars, beta, cov, cells=ncel[3], dics=X[3],
                    m.prior="normal", d.prior="uniform",
```

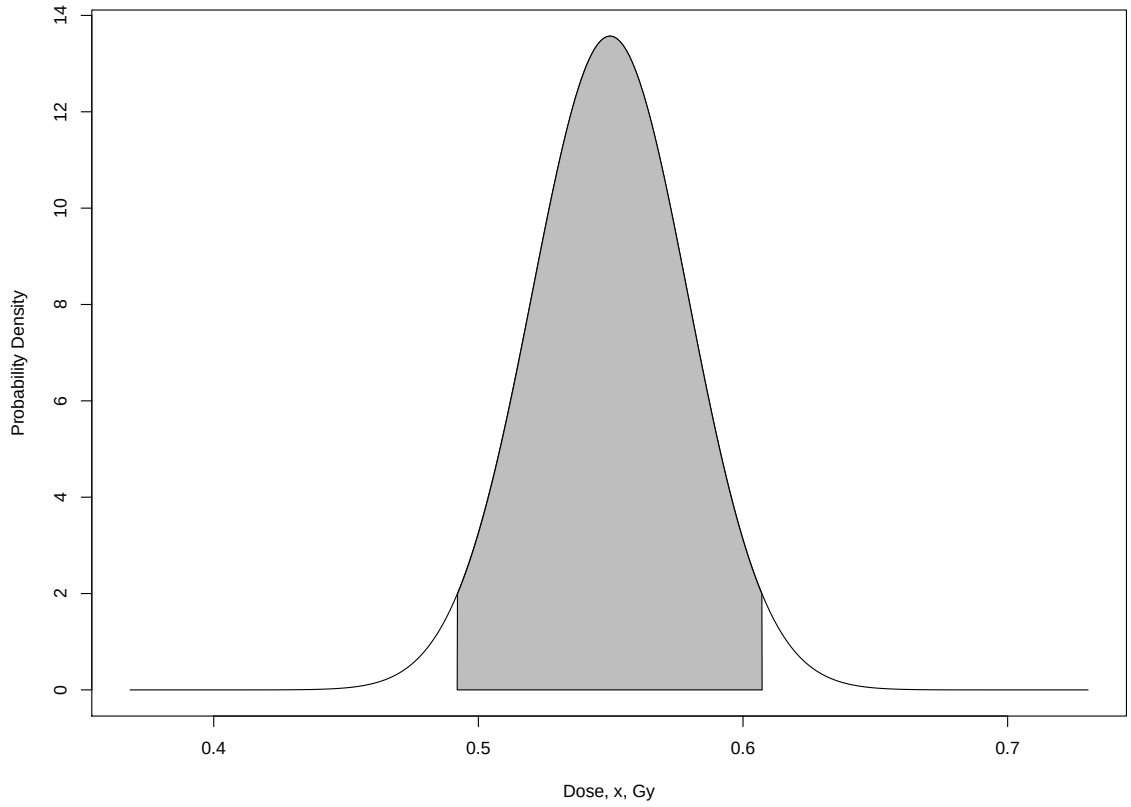



Fig. 2.3 95% HPD interval of the calibrative density of the 0.5Gy test data for a normal mean prior and a $U(0, \infty)$ dose prior.

```
                    prior.param=c(0, 10))
plot(ex1, ci=T, cr=0.95)
```

where, if necessary, one can also specify a gamma prior as input for the mean and dose priors. We see clearly that the 95% interval just slightly covers the 0.5Gy dose (not encompassed by a 90% interval), producing an expected dose of 0.55Gy. Although this remains a reasonable estimate, particularly in the circumstance of needing a quick value, the package only takes into account the total number of cells as well as the total aberration count and therefore no information of the sample's distribution (i.e. many foci-free or single-foci cells, present overdispersion?). Additionally, this approach would not be suitable for calibration data consisting of only 3 design doses with a single sample per dose, as is the case with our H2AX datasets, since it becomes meaningless to re-fit a calibration curve with only 2 dose points.

The assessment of the absorbed dose is particularly reliable in cases of acute, uniform and whole-body exposures. However, the scenarios of most radiation accidents result in partial-body exposures or non-uniform dose distribution, leading to a differential exposure of lymphocytes in the body. Subsequently for the exposed individuals, their

blood will contain a mixture of cells showing no radiation impact at all and cells featuring a distribution of counts according to the dose of exposure. As a consequence, the produced yield of dicentric or foci aberrations in a patient blood sample will become overdispersed and therefore no longer conform to the Poisson distribution. For this reason, the uncertainty methods discussed in Section 2.4 for the purpose of whole-body exposures will need to be updated to account for this overdispersion.

In the context of dicentrics, there do exist some methods to infer the degree of partial body exposure from the overdispersion. However, it remains in such exposure scenarios that there are no statistical procedures to follow for the $\gamma$-H2AX assay. Part of the challenge resides in pinpointing the exact cause(s) of the overdispersion. For instance, experimental factors which generate variability in the scoring process of cells are all absorbed by the dispersion value. Such dispersion-generating effects are generally present for both manual and automated scoring of $\gamma$-H2AX foci. To motivate the developments which are to come, in the following chapter we begin by defining explicitly what dispersion means for biomarker data before understanding its relevance to the Poisson mean-variance assumption and its behaviour across various levels of dose and exposure.

# Chapter 3

# Dispersion in count data

The goal in biological dosimetry is to estimate the dose of ionising radiation absorbed by an exposed individual by using chromosome damage in peripheral lymphocytes as a biomarker of exposure. The radiation dose that an overexposed individual has received is estimated by means of a dose response calibration curve which is created by exposures of human blood cells to different and appropriate doses of radiation. In addition, it is equally desirable to quantify the fraction of exposure at which the dose has been delivered. In either case, the endpoint of interest is the number of aberrations (in this context dicentrics or H2AX foci) observed. It has been reported in the literature that the foci distribution in scored blood cells becomes overdispersed in both WB (whole-body) and PB (partial-body) irradiation scenarios. In this chapter, we attempt to explain the significance of a dispersion index (the ratio of the sample variance to the mean) and its magnitude in relation to the detection of inhomogeneous exposures.

## 3.1 What is dispersion?

Typically when we speak of dispersion, we are referring to the variance divided by the mean. More precisely, and ignoring (for now) the presence of covariates, dispersion is defined by

$$\phi = \frac{\mathrm{Var}(y_{ij})}{E(y_{ij})} \tag{3.1.1}$$

which can be estimated through the *dispersion index*

$$\hat{\delta} = \frac{1}{\bar{y}} \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}{N - 1},$$

where $N = \sum_{i=1}^{k} n_i$ is the total number of observed counts, and $\bar{y} = N^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$ is their overall mean. Similarly, the slide-wise dispersion index can be computed via

$$\hat{\delta}_i = \frac{1}{y_i} \frac{\sum_{j=1}^{n_i} (y_{ij} - y_i)^2}{n_i - 1}.$$

If $\hat{\delta} > 1$ one speaks of overdispersion, while for $\hat{\delta} < 1$ one has underdispersion. Underdispersed data are unusual in the field of biodosimetry, and their observation is sometimes indicative that something during the experiment did not work, or perhaps due to other mechanisms currently not very well-known [83]. There are several possible factors which may contribute to overdispersion. A certain role is played by technical variations, such as in the intensity filter used for the foci scoring. Specifically, for low foci rates the semi-automated imaging software which aids the foci scoring tends to produce spurious foci by over-enhancing background signals. Other sources of overdispersion may relate to physical issues with the slides, issues with the radiation source itself or the placement of the samples, issues relating to the antibodies used to produce foci, the microscope, and the scorer. Furthermore, it is likely to assume a 'learning effect' for the scorer who may be tempted to discard samples which do not fit the previously observed pattern. In some circumstances, the cause of the overdispersion may be apparent from the nature of the data collection process. In most cases, however, it is often difficult to infer the precise cause leading to the overdispersion.

## 3.2 Testing for overdispersion

Early evidence has suggested that the distribution of $\gamma$-H2AX foci among the scored blood cells can be analysed by employing the same methods used for the dicentric assay [52, 85]. According to the manual of the IAEA, the standard procedure to detect PBI uses the well-known Papworth $U$-test to determine whether the sample dispersion index is significantly different from 1. The test statistic remains

$$U = \left(\hat{\delta} - 1\right) \sqrt{\frac{N - 1}{2\left(1 - \frac{1}{S}\right)}}$$

where $S = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = N\bar{y}$ or for some slide $i$,

$$U_i = \left(\hat{\delta}_i - 1\right) \sqrt{\frac{n_i - 1}{2\left(1 - \frac{1}{s_i}\right)}}.$$

in which $s_i = \sum_{j=1}^{n_i} y_{ij} = n_i y_i$.

For testing at the 5% significance level, values of $U$ greater than 1.64 indicate rejection of the null hypothesis $H_0 : \phi = 1$ in preference for $H_1 : \phi > 1$. We note that

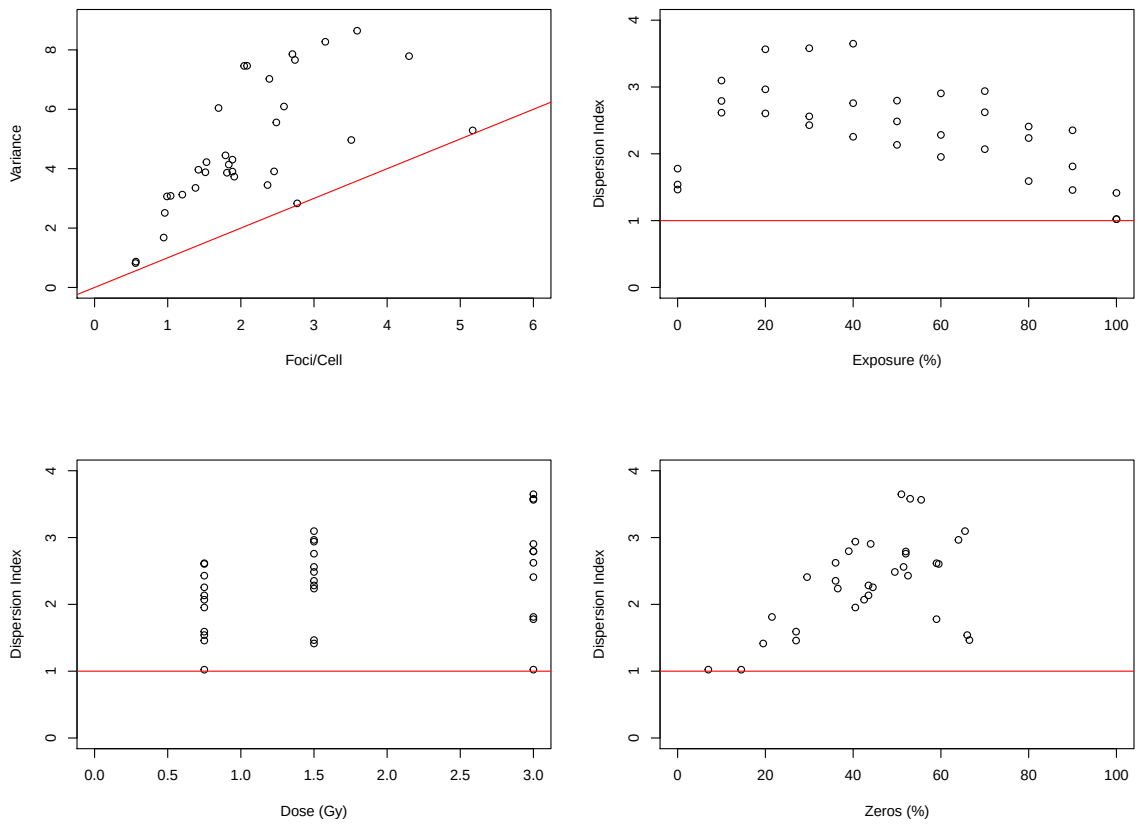



Fig. 3.1 Dispersion index behaviour for PHE-Foci1 4h data against exposure, dose and proportion of zeros. Clearly, as one reaches a higher level of dose, the number of foci tends to increase, yielding a reduced percentage of zero counts.

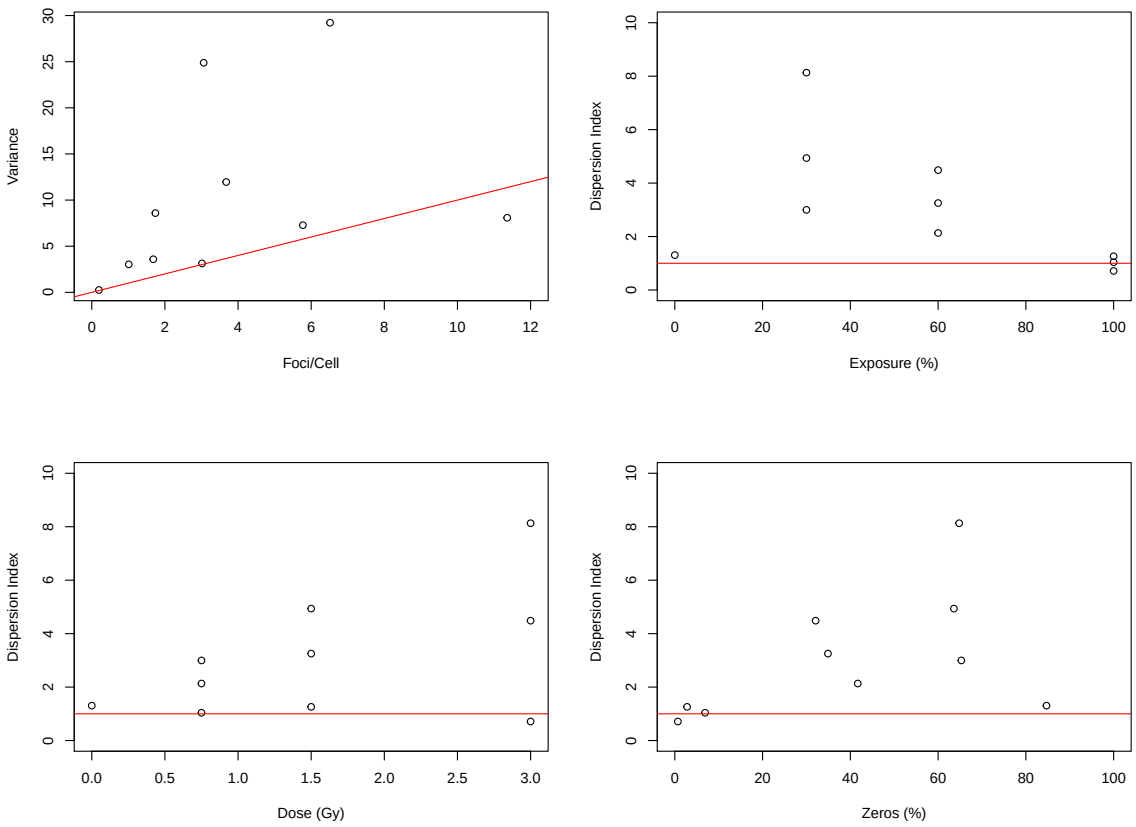



Fig. 3.2 Dispersion index behaviour for PHE-Foci2 dataset.

this test is usually only applied in the context of cytogenetic biomarker data, in which any significant overdispersion would normally be used as evidence for the presence of PBI. It is clear from Figures 3.1 and 3.2 that the distribution of H2AX foci can become overdispersed in the case of both PBI and WBI, noting that the majority of data points do not follow the unity line (i.e. $\hat{\delta}_i > 1$ in the top-left panels). In the next section we compare the magnitude of $U$ for H2AX and dicentric data based on similar levels of exposure.

## 3.3 Detecting zero-inflation

It can be speculated from the bottom-right panels of Figures 3.1 and 3.2 that data which exhibit a high proportion of aberration-free cells tend to have an increased variance and therefore a larger dispersion. It is desirable to clarify whether the data are zero-inflated or not, that is, to check the proportions for excess zeros. It is known that zero-inflation can lead to the rejection of the Poisson hypothesis, in cases where this is not rejected when just the $U$-test is applied. Accordingly, it is suggested to use an exact zero-inflation test.

The CR-test of goodness-of-fit for the Poisson distribution was firstly presented by Rao and Chakravarti on the basis of a problem related to occupancy distributions [84]. The CR-test to contrast the null hypothesis $H_0$: Data are Poisson distributed, against the alternative $H_1$: Data are zero-inflated consists of calculating a p-value $P(N_0 \geq n_0)$, where here $N_0$ is the random variable representing the number of aberration-free cells. Concretely, the p-value for some slide $i$ is computed as follows [34]

$$\begin{aligned} P(N_0 \geq n_0) &= \sum_{h=n_0}^{n_i} \sum_{l=h}^{n_i} (-1)^{l-h} \binom{n_i}{l} \binom{l}{h} \left(1 - \frac{l}{n_i}\right)^{s_i} \\ &= \sum_{h=n_0}^{n_i} \frac{(-1)^{h-n_0}}{(n_0-1)!(h-n_0)!} \binom{n_i}{h} \left(1 - \frac{h}{n_i}\right)^{s_i} \end{aligned} \tag{3.3.1}$$

where it is assumed that foci are randomly distributed between cells with the same probability $1/n_i$. For dealing with large values of $n_i$ and $s_i$, [84, 34] suggests to use an asymptotic approximation of (3.3.1) based on a normalised $N_0$. It follows that

$$P(N_0 > n_0) \approx 1 - \Phi(z_0), \tag{3.3.2}$$

where $\Phi(z_0)$ is the CDF of the standard normal evaluated at $z_0 = (n_0 - \mathrm{E}(N_0))/\sqrt{\mathrm{Var}(N_0)}$. The expectation and variance of $N_0$ can be computed directly through

$$\mathrm{E}(N_0) = n_i \left(\frac{n_i - 1}{n_i}\right)^{s_i},$$

$$\mathrm{Var}(N_0) = n_i(n_i - 1)\left(\frac{n_i - 2}{n_i}\right)^{s_i} + n_i\left(\frac{n_i - 1}{n_i}\right)^{s_i}\left(1 - n_i\left(\frac{n_i - 1}{n_i}\right)^{s_i}\right).$$

The reason why the CR-test is particularly suitable to the Poisson distribution is because it allows us to evaluate the problem of zero-inflation in the data, which cannot be identified using only the $U$-test. For measuring the degree of zero-inflation, however, Puig and Valero [82] proposed to use a zero-inflation index $zi$. In the case of a single slide, it is defined as

$$zi = 1 + \frac{\log\left(\frac{n_0}{n_i}\right)}{y_i}. \tag{3.3.3}$$

If the sample is Poisson distributed then one has $e^{-y_i} = n_0/n_i$ or $zi = 0$, with $zi > 0$ meaning it is zero-inflated. It is clear that both $N_0$ and the associated index $zi$ seem to be suitable measures for exploring whether or not data stem from a Poisson distribution.

Tables 3.1 and 3.2 show the results obtained using the $U$-test and the CR-test for both PBI and WBI scenarios for H2AX scored datasets. Accordingly, both tests are one-tailed since we are interested in testing a Poisson distribution against overdispersed or zero-inflated distributions, respectively. Reasonably, for a one-side $U$-test at the 5% significance level, the null hypothesis is rejected when $U \geq 1.64$. Firstly for the PHE-Foci1 dataset, data seem to be overdispersed with the exception of strong evidence for equidispersion (p-values $>>$ 0.05) in the 0.75Gy and 3Gy WB samples. However, both the p-values of the CR-test and $zi$ indices indicate all samples are hugely zero-inflated. This is also evident by comparing the difference in proportions between the actual zeros and the expected zeros under a Poisson (i.e. $n_0/n_i >> e^{-y_i}$). Furthermore, this suggests that the apparent overdispersion is mostly due to observed excess foci-free cells.

Upon drawing particular attention to the 100% exposure data, it would appear that the foci distribution adheres well to a Poisson (albeit some evidence in favour of rejecting $H_0$ for the 1.5Gy sample). The Poisson assumption of equidispersion means under $H_0$ we would expect

$$\hat{\delta} \sim \frac{\chi^2_{n_i-1}}{n_i - 1}. \tag{3.3.4}$$

Again, the null hypothesis is accepted only for the 0.75Gy and 3Gy WB samples when comparing $\hat{\delta}$ with the critical value $\frac{\chi^2_{0.95,199}}{199} = 1.170$. In summing over $k$ slides, one has

| Exposure (%) | Dose (Gy) | $n_0/n_i$ | $s_i$ | $e^{-y_i}$ | $\hat{\delta}$ | $U$ | CR p-value | $zi$ |
|---|---|---|---|---|---|---|---|---|
| | 0.75 | 0.525 | 276 | 0.252 | 2.430 | 14.283 | $< 0.001$ | 0.533 |
| 30 | 1.5 | 0.515 | 303 | 0.220 | 2.561 | 15.592 | $< 0.001$ | 0.562 |
| | 3 | 0.530 | 417 | 0.124 | 3.580 | 25.770 | $< 0.001$ | 0.696 |
| | 0.75 | 0.445 | 367 | 0.160 | 2.255 | 12.539 | $< 0.001$ | 0.559 |
| 40 | 1.5 | 0.520 | 306 | 0.217 | 2.758 | 17.568 | $< 0.001$ | 0.573 |
| | 3 | 0.510 | 409 | 0.129 | 3.648 | 26.447 | $< 0.001$ | 0.671 |
| | 0.75 | 0.405 | 382 | 0.148 | 1.953 | 9.520 | $< 0.001$ | 0.527 |
| 60 | 1.5 | 0.435 | 377 | 0.152 | 2.283 | 12.814 | $< 0.001$ | 0.558 |
| | 3 | 0.440 | 541 | 0.067 | 2.905 | 19.017 | $< 0.001$ | 0.696 |
| | 0.75 | 0.270 | 491 | 0.086 | 1.592 | 5.907 | $< 0.001$ | 0.467 |
| 80 | 1.5 | 0.365 | 497 | 0.083 | 2.236 | 12.346 | $< 0.001$ | 0.594 |
| | 3 | 0.295 | 718 | 0.028 | 2.408 | 14.056 | $< 0.001$ | 0.660 |
| | 0.75 | 0.145 | 554 | 0.062 | 1.022 | 0.221* | $< 0.001$ | 0.303 |
| 100 | | | | | | (0.413) | | |
| | 1.5 | 0.195 | 702 | 0.030 | 1.414 | 4.137 | $< 0.001$ | 0.534 |
| | 3 | 0.070 | 1034 | 0.006 | 1.023 | 0.227* | $< 0.001$ | 0.486 |
| | | | | | | (0.410) | | |

* Indicates the $U$ test statistic is non-significant at the 5% significance level.

Table 3.1 Results from PHE-Foci1 dataset. P-values for estimates supporting a Poisson distribution are given in paranthesis.

$$\hat{\delta} \sim \frac{\chi^2_{N-k}}{N-k}. \tag{3.3.5}$$

Initial consideration of the merged samples (i.e. calibration data) reveals a non-Poisson nature ($\hat{\delta} = 2.724$, $U = 29.847$) and for $k = 3$ in (3.3.5), this leads to supporting the decision made from the $U$-test to reject $H_0$ ($\hat{\delta} = 2.724 > \frac{\chi^2_{0.95,597}}{597} = 1.097$).

The results presented in Table 3.2 for the PHE-Foci2 dataset are more complex to deal with. For the 3Gy/WB sample, we detect the rare eventuality of underdispersion ($U \leq -1.64$). Although the $zi$ index and p-value confirm zero-inflation, this is mostly a consequence of higher counts producing a larger mean and hence it is likely that there is an alternate source/mechanism contributing to the foci distribution of this particular sample. We can only assume this has occured due to a saturation (3Gy being a high dose) of what could be scored i.e. too many (and/or overlapping) foci in such a small cell - lymphocytes combined with some caveats (e.g. magnification, staining quality etc.). From the zero proportions, we examine a surprising 14 foci-free cells in the PHE-Foci1 3Gy/WB sample but half this amount from the PHE-Foci2 sample. For WBI, and particularly for large doses, it is expected that the number of zeros should be minimal and on this basis it would seem initially that the results from the PHE-Foci2 provide a more accurate representation (despite the unusual underdispersion).

| Exposure (%) | Dose (Gy) | $n_0/n_i$ | $s_i$ | $e^{-y_i}$ | $\hat{\delta}$ | $U$ | $1-\Phi(z_0)$ | $zi$ |
|---|---|---|---|---|---|---|---|---|
| 30 | 0.75 | 0.653 | 1012 | 0.363 | 2.997 | 44.652 | $< 0.001$ | 0.579 |
| | 1.5 | 0.636 | 1740 | 0.176 | 4.936 | 87.997 | $< 0.001$ | 0.740 |
| | 3 | 0.648 | 3060 | 0.047 | 8.131 | 159.40 | $< 0.001$ | 0.858 |
| 60 | 0.75 | 0.417 | 1682 | 0.186 | 2.133 | 25.339 | $< 0.001$ | 0.480 |
| | 1.5 | 0.349 | 3676 | 0.025 | 3.253 | 50.353 | $< 0.001$ | 0.714 |
| | 3 | 0.321 | 6516 | 0.001 | 4.484 | 77.882 | $< 0.001$ | 0.826 |
| 100 | 0.75 | 0.069 | 3014 | 0.049 | 1.039 | 0.882 (0.189) | $< 0.001$ | 0.113 |
| | 1.5 | 0.028 | 5773 | 0.003 | 1.261 | 5.837 | $< 0.001$ | 0.381 |
| | 3 | 0.007 | 11360 | $<0.001$ | 0.712 | -6.444 | $< 0.001$ | 0.563 |

Table 3.2 Results from PHE-Foci2 dataset.

| Exposure (%) | Dose (Gy) | $n_0/n_i$ | $s_i$ | $e^{-y_i}$ | $\hat{\delta}_i$ | $U$ | CR p-value | $zi$ |
|---|---|---|---|---|---|---|---|---|
| 50 | 0.5 | 0.875 | 169 | 0.845 | 1.436 | 9.773 | $< 0.001$ | 0.210 |
| | 0.7 | 0.849 | 167 | 0.812 | 1.548 | 10.980 | $< 0.001$ | 0.214 |
| | 1 | 0.800 | 188 | 0.731 | 1.711 | 12.334 | $< 0.001$ | 0.288 |
| 75 | 0.5 | 0.791 | 231 | 0.749 | 1.432 | 8.646 | $< 0.001$ | 0.189 |
| | 0.7 | 0.758 | 203 | 0.713 | 1.353 | 6.134 | $< 0.001$ | 0.182 |
| | 1 | 0.658 | 203 | 0.602 | 1.383 | 5.418 | $< 0.001$ | 0.174 |
| 100 | 0.5 | 0.710 | 319 | 0.671 | 1.286 | 5.720 | $< 0.001$ | 0.141 |
| | 0.7 | 0.593 | 343 | 0.565 | 1.182 | 3.163 | 0.002 | 0.086 |
| | 1 | 0.423 | 372 | 0.395 | 1.153 | 2.171 | 0.032 | 0.074 |

Table 3.3 Results from PHE-Dicentric dataset.

On noting that the samples produced for the PHE-Foci2 dataset are five times larger than those in the PHE-Foci1 dataset, we would anticipate the PHE-Foci2 data to be less error-prone and more reliable. We were informed by PHE that the PHE-Foci2 samples were from a more experienced scorer, as compared to the scorer who conducted the experiments of the PHE-Foci1 data. Since an inexperienced scorer will tend to want to see cells similar to each other, this can have the impact of under-reporting foci (cells with more damage are more difficult to score) thereby reducing the dispersion towards 1. Only through experience will a scorer be able to recognise patterns as well as distinguish between innate and radiation-induced variability.

On further comparison of the same exposure fractions we identify that the estimates of $\hat{\delta}$ are larger in all the overdispersed PHE-Foci2 samples except the 1.5Gy/WB sample. Based on the critical value 1.075, computed from (3.3.4), it remains that $H_0$ is retained only for the 0.75Gy/ and 3Gy/WB samples. For the combined WB data, however, one has $U = 66.33$ and $\hat{\delta} = 2.713 > \frac{\chi^2_{0.95,2997}}{2997}$ thus a Poisson fit would be deemed unsuitable.

As a result of the increased dispersions, one can also deduce a positive correlation between dose level and $U$-value in the partially-irradiated samples, thus the higher the dose the further that sample's distribution deviates from a Poisson.

The results for the PHE-Dicentric dataset are provided in Table 3.3, where immediately we identify that the values of $\hat{\delta}_i$, $U$ and $zi$ for the partially-irradiated samples are all smaller in contrast to the H2AX datasets. For WBI, however, the observed variance-mean ratios are larger than expected (all samples being significant), with the merged data also rejecting $H_0$ ($U = 5.497$ and $\hat{\delta} = 1.286 > \frac{\chi^2_{0.95,1798}}{1798} = 1.055$). Since the $n_0/n_i$ are relatively close to $e^{-y_i}$ then for this dataset, zero-inflation is most likely not the main contributor to the overdispersion.

# Chapter 4

# Modelling overdispersed radiation biomarkers

While for the dicentric assay it is assumed that any significant overdispersion would be used as evidence for the presence of partial exposure, this would be incorrect in our context since we have seen that the Poisson distribution of dicentrics can become overdispersed in the case of whole-body exposure. A suitable way to deal with such overdispersion for count data is based on the generalised linear model framework outlined in Section 2.1, where the most common approach is a "quasi-likelihood", with Poisson-like assumptions (which we refer to as a quasi-Poisson from hereafter) [104] or alternatively through a negative binomial model. In this chapter we intend to summarise the statistical and conceptual basics of the quasi-Poisson and negative binomial models, including the estimation of the dispersion parameter as well as their variance.

## 4.1 Quasi-Poisson

The violation of the mean-variance assumption can be described by generalising (2.2.7),

$$\frac{\mathrm{Var}(y_{ij}|x_i)}{E(y_{ij}|x_i)} = \phi \tag{4.1.1}$$

for some constant dispersion, $\phi > 0$, which is also the 'dispersion parameter' which we refer to in (2.1.2). If $\phi > 1$ one speaks of overdispersion, while for $\phi < 1$ one has underdispersion. Poisson regression models can be easily adapted to allow for situation (4.1.1), since the dispersion cancels out from the score equations and so the estimates

of regression parameters are unaffected (i.e. one obtains identical calibration curves, for example to those shown in Figures 2.1 and 2.2). One speaks then of *quasi-Poisson* (QP) regression models, which have gained some interest specifically in the field of biodosimetry [28].

Under the quasi-Poisson model, the dispersion parameter presented in (4.1.1) can be estimated by equating the Pearson $X^2$ goodness-of-fit statistic to the residual degrees of freedom $\nu = N - w$, that is

$$\hat{\phi} = \frac{X^2}{\nu} = \frac{1}{\nu}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(y_{ij} - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{4.1.2}$$

where $w$ is the number of model parameters and $\hat{\mu}_i = g^{-1}(x_i'\hat{\boldsymbol{\beta}})$. McCullagh and Nelder [71] discuss the advantage of basing the estimation of the dispersion parameter on (4.1.2) as opposed to using the residual deviance. The standard errors associated with the estimated coefficients, $\hat{\boldsymbol{\beta}}$, will be the same as for the non-dispersed Poisson model but instead inflated by the factor $\sqrt{\hat{\phi}}$.

A key question is how does dispersion behave under aggregation? For the aggregated counts, one has

$$\begin{aligned}\mathrm{Var}(s_i|x_i) &= \mathrm{Var}\left(\sum_{j=1}^{n_i} y_{ij}|x_i\right) \overset{*}{=} \sum_{j=1}^{n_i}\mathrm{Var}(y_{ij}|x_i) \\ &= \sum_{j=1}^{n_i}\phi E(y_{ij}|x_i) = \phi\sum_{j=1}^{n_i}E(y_{ij}|x_i) = \phi E(s_i|x_i),\end{aligned} \tag{4.1.3}$$

where the step (*) is a consequence of conditional independence assumption; so once again $\mathrm{Var}(s_i|x_i)/E(s_i|x_i) = \phi$ so that the dispersion is, theoretically, invariant to aggregation. In other words, we assume that the true dispersion, $\phi$, is indeed the same for the raw and aggregated data. For aggregated counts $s_i = \sum_{j=1}^{n_i} y_{ij}$, $i = 1, \ldots, k$ (equivalently expressed through the yields $y_i = s_i/n_i$), with aggregated data model as specified in (2.2.8), the value of the dispersion can be estimated by

$$\hat{\phi}_{agg} = \frac{1}{\nu}\sum_{i=1}^{k}\frac{(s_i - n_i\hat{\mu}_i)^2}{n_i\hat{\mu}_i} = \frac{1}{\nu}\sum_{i=1}^{k} n_i\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, \tag{4.1.4}$$

where here $\nu = k - w$ and $\hat{\mu}_i$ is as above.

### 4.1.1 Dispersion variability

If indeed our fitted model is assumed to be correct then we would expect $X^2/\phi$ to have a $\chi^2_\nu$ distribution implying that

$$E(X^2) = \phi\nu$$

and

$$\mathrm{Var}(X^2) = 2\phi^2\nu\,.$$

It follows that

$$\begin{aligned} E(\hat{\phi}_\nu) &= \frac{1}{\nu}E(X^2) \\ &= \phi, \end{aligned} \tag{4.1.5}$$

so both the raw and aggregated dispersion estimate are unbiased, and the variance of $\hat{\phi}$ is given by

$$\begin{aligned} \mathrm{Var}(\hat{\phi}_\nu) &= \frac{1}{\nu^2}\mathrm{Var}(X^2) \\ &= \frac{2\phi^2}{\nu}. \end{aligned} \tag{4.1.6}$$

Provided that one has a suitable estimate for $\phi$, (4.1.6) can be used to obtain an estimate for the dispersion variance.

While (4.1.1) is already a considerable generalisation of the 'plain' Poisson model, it should be said for completeness that in practice the dispersion may depend on covariates $x_i$, that is $\phi = \phi(x_i)$. Covariate-dependent dispersion cannot be expressed by quasi-Poisson regression and requires fitting more advanced models. Several models have been developed for this purpose which include the generalised Poisson [23], Hermite [44], and more recently COM-Poisson [20]. In this work, we dedicate our attention to the negative binomial and zero-inflated models.

### 4.1.2 Dispersion confidence interval

It is possible to construct a $100(1-\epsilon)\%$ confidence interval for the dispersion parameter $\phi$. Assuming that

$$\frac{X^2}{\phi} \sim \chi^2_\nu,$$

then one has

$$
\begin{aligned}
1-\epsilon &= \mathcal{P}\left[\chi^2_{\nu,1-\epsilon/2} \geq \frac{X^2}{\phi} \geq \chi^2_{\nu,\epsilon/2}\right] \\
&= \mathcal{P}\left[\frac{1}{\chi^2_{\nu,\epsilon/2}} \geq \frac{\phi}{X^2} \geq \frac{1}{\chi^2_{\nu,1-\epsilon/2}}\right] \\
&= \mathcal{P}\left[\frac{X^2}{\chi^2_{\nu,\epsilon/2}} \geq \phi \geq \frac{X^2}{\chi^2_{\nu,1-\epsilon/2}}\right].
\end{aligned}
$$

Therefore one can obtain separate confidence intervals for the raw and aggregated dispersion using the following:

$$
\begin{aligned}
\frac{X^2}{\chi^2_{N-w,\epsilon/2}} \geq \phi \geq \frac{X^2}{\chi^2_{N-w,1-\epsilon/2}} \\
\frac{X^2}{\chi^2_{k-w,\epsilon/2}} \geq \phi_{agg} \geq \frac{X^2}{\chi^2_{k-w,1-\epsilon/2}}.
\end{aligned}
$$

## 4.2 Negative binomial regression

The quasi-Poisson estimation handles overdispersion by moving away from a complete distributional specification. Alternatively, it is possible to define a distribution that permits more flexible modelling of the variance than the Poisson. The standard parametric model to account for overdispersion is the negative binomial (NB) whose probability mass function is defined by

$$
P(Y_{ij} = y_{ij}|\mu_i, \alpha) = \frac{\Gamma(y_{ij} + \frac{\mu_i^{1-c}}{\alpha})}{y_{ij}!\Gamma(\frac{\mu_i^{1-c}}{\alpha})}(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}(1+\frac{\mu_i^{-c}}{\alpha})^{-y_{ij}} \tag{4.2.1}
$$

where $\alpha > 0$ is an overdispersion parameter and the index $c \in (0,1)$ identifies the form of the underlying NB distribution. An estimated $c$ can be obtained using maximum likelihood estimation [48], however, here we consider only the models given by $c = 0$ and 1. For $c = 0$, we have a linear-variance NB regression, with $\text{Var}(Y_{ij}) = \mu_i(1+\alpha)$, denoted by NB1. Taking $c = 1$ gives the more usual quadratic-variance NB model, with $\text{Var}(Y_{ij}) = \mu_i(1+\mu_i\alpha)$, which is denoted by NB2. For both models, we will continue to model $\mu_i$ through (2.2.2) using the identity link function. We conclude this subsection by noting that the overdispersion in the NB1 case is the multiplicative factor $1+\alpha$ which does not depend on $\mu_i$, unlike the NB2. This is particularly convenient in our context since it is desirable to relate the irradiated fraction to a single constant/value which is independent of covariates such as dose. The NB1 dispersion can be viewed as an alternative to the quasi-Poisson, replacing $\phi$ with $1+\alpha$ in (4.1.1). In contrast,

despite both the NB1 and NB2 not being members of the exponential family, they have the distinct advantage that their likelihood function can be formally defined.

### 4.2.1 NB MLE

For observations $y_{ij} \sim \text{NB}(\mu_i, \alpha)$, the general NB log-likelihood function, $\ell$, can be written as

$$\begin{aligned}\ell(\mu_i, \alpha) = &\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\sum_{m=0}^{y_{ij}-1}\ln\left(m + \frac{\mu_i^{1-c}}{\alpha}\right)\right) - \ln(y_{ij}!) \\ &- \frac{\mu_i^{1-c}}{\alpha}(1+\alpha\mu_i^c) - y_{ij}\ln\left(1+\frac{\mu_i^{-c}}{\alpha}\right).\end{aligned}$$

For the gamma function, $\Gamma(.)$, we make use of the property $\Gamma(a+b)/\Gamma(b) = \prod_{m=0}^{a-1} m+b$, assuming $a$ and $m$ are integers [14]. The maximum likelihood estimates $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ are then the solutions to the following first-order conditions

$$\begin{aligned}&\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\sum_{m=0}^{y_{ij}-1}\frac{1-c}{\alpha\mu_i^c+\mu_i}\right) - \frac{(1-c)\ln(1+\alpha\mu_i^c)}{\alpha\mu_i^c} - \frac{c}{1+\alpha\mu_i^c} = 0 \\ &\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(-\sum_{m=0}^{y_{ij}-1}\frac{\mu_i}{m\alpha^2\mu_i^c+\alpha\mu_i}\right) + \frac{\mu_i((\mu_i^{-c}+\alpha)\ln(1+\alpha\mu_i^c)-\alpha)}{\alpha^2(1+\alpha\mu_i^c}) = 0.\end{aligned} \tag{4.2.2}$$

For use with the construction of confidence intervals for $\alpha$ and $\boldsymbol{\beta}$, the model output will implicitly produce standard errors ($\text{SE}(\hat{\alpha})$ and $\text{SE}(\hat{\boldsymbol{\beta}})$) through the Fisher information matrix. Furthermore, given these quantities, a 95% confidence interval for the dispersion $\phi$ can be found through

$$\hat{\phi} \pm z_{0.975}\text{SE}(\hat{\phi}) \tag{4.2.3}$$

where $\hat{\phi} = 1 + \hat{\alpha}$ for the NB1 or $\hat{\phi}(\hat{\mu}_i) = 1 + \hat{\alpha}\hat{\mu}_i$ for the NB2 and $z_{0.975}$ is the 97.5th quantile of the standard normal distribution (1.96).

## 4.3 Zero-inflated models

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. Count datasets with excessive zeros are extensive in a wide variety of disciplines, such as public health and environmental science. To account for a preponderance of zero counts, zero-inflated models can be used which describe the data as a combination of two distributions: a distribution which takes a single value at zero and a count distribution such as the Poisson or NB. These models are particularly useful in the context of partial-body irradiation scenarios, which feature a mixture of populations of non-irradiated and irradiated cells.

### 4.3.1 Zero-inflated Poisson

The zero-inflated Poisson (ZIP) regression model was first introduced by Lambert [62] who applied the model to data gathered from a quality control study. After which, the ZIP regression model has been well-studied in the literature [15, 22]. The probability mass function for the ZIP model is defined as:

$$P(Y_{ij} = y_{ij}|\mu_i, p_i) = \begin{cases} p_i + (1-p_i)e^{-\mu_i}, & \text{for } y_{ij} = 0 \\ (1-p_i)\frac{e^{-\mu_i}\mu_i^{y_{ij}}}{y_{ij}!}, & \text{for } y_{ij} > 0 \end{cases} \tag{4.3.1}$$

where $0 \leq p_i \leq 1$ and $\mu_i > 0$, possibly depending on covariates such as dose. Here, $\mu_i$ refers to the mean of the underlying Poisson distribution and $p_i$ is the zero-inflation parameter. The ZIP model has the properties: $E(y_{ij}|x_i) = (1-p_i)\mu_i = \lambda_i$ and $\text{Var}(y_{ij}|x_i) = (1-p_i)\mu_i(1+p_i\mu_i)$ and reduces to a Poisson when $p_i = 0$. Following the notation in (4.1.1), the ZIP dispersion is given by

$$\frac{\text{Var}(y_{ij}|x_i)}{E(y_{ij}|x_i)} = \frac{(1-p_i)\mu_i(1+p_i\mu_i)}{(1-p_i)\mu_i} = 1 + p_i\mu_i. \tag{4.3.2}$$

A ZIP model takes into account that the zero observations have two different origins: zeros which are produced at random by the Poisson distribution, while some others, with proportion $p_i$, are considered *structural*. The structural zeros are dependent on the nature of data, in our case by non-irradiated lymphocytes following partial-body exposure.

The ZIP log-likelihood function is

$$\ell(\mu_i, p_i) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(I_{(y_{ij}=0)}\ln\left[p_i + (1-p_i)e^{-\mu_i}\right] + I_{(y_{ij}>0)}\ln\left[(1-p_i)\frac{e^{-\mu_i}\mu_i^{y_{ij}}}{y_{ij}!}\right]\right)$$

with corresponding score equations given by

$$\begin{aligned}\frac{\partial \ell}{\partial p_i} &= \sum_{j=1}^{n_i}\left(I_{(y_{ij}=0)}\left[\frac{1-e^{-\mu_i}}{p_i+(1-p_i)e^{-\mu_i}}\right]+I_{(y_{ij}>0)}\left[\frac{-1}{1-p_i}\right]\right)\\ &= \frac{N_0(1-e^{-\mu_i})}{p_i+(1-p_i)e^{-\mu_i}}-\frac{N-N_0}{1-p_i};\end{aligned} \tag{4.3.3}$$

$$\begin{aligned}\frac{\partial \ell}{\partial \mu_i} &= \sum_{j=1}^{n_i}\left(I_{(y_{ij}=0)}\left[-\frac{(1-p_i)e^{-\mu_i}}{p_i+(1-p_i)e^{-\mu_i}}\right]+I_{(y_{ij}>0)}\left[-1+\frac{y_{ij}}{\mu_i}\right]\right)\\ &= -\frac{N_0(1-p_i)e^{-\mu_i}}{p_i+(1-p_i)e^{-\mu_i}}-(N-N_0)+\frac{S}{\mu_i}.\end{aligned} \tag{4.3.4}$$

By equating (4.3.3) to zero, the MLE of $p_i$ is found to be

$$\hat{p}_i = \frac{N_0 - Ne^{-\hat{\mu}_i}}{N\left(1-e^{-\hat{\mu}_i}\right)}. \tag{4.3.5}$$

It is acceptable to model the mean of the corresponding zero-inflated distribution, $\lambda_i$, via the linear predictor in equation (2.2.3) as opposed to modelling the mean of the underlying Poisson distribution, $\mu_i$, which are related via

$$\mu_i = \frac{\lambda_i}{1-p_i}. \tag{4.3.6}$$

By substituting either (4.3.6) into (4.3.5) or (4.3.5) into (4.3.4) set to zero, the MLE of $\mu_i$ is the solution to the expression:

$$\hat{\mu}_i(N-N_0) - S\left(1-e^{-\hat{\mu}_i}\right) = 0. \tag{4.3.7}$$

From the ZIP property $\mathrm{Var}(y_{ij}) \geq E(y_{ij}) = \lambda_i$, zero-inflation can be seen as a special form of overdispersion. However, the non-zero part of the foci-count distribution may still be overdispersed even after accounting for zero-inflation. For this reason, for data which stem from full- or partial-body exposure, it is sensible to consider overdispersion and zero–inflation as two separately identifiable model properties.

### 4.3.2 Zero-inflated NB

When overdispersion is both due to data heterogeneity and the excess of zeros, the zero-inflated negative binomial (ZINB) regression model proposed by [38] is often more appropriate than the ZIP. The ZINB probability mass function is given by

$$P(Y_{ij} = y_{ij}|\mu_i, p_i, \alpha) = \begin{cases} p_i + (1-p_i)(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}, & \text{for } y_{ij} = 0 \\ (1-p_i)\frac{\Gamma(y_{ij}+\frac{\mu_i^{1-c}}{\alpha})}{y_{ij}!\Gamma(\frac{\mu_i^{1-c}}{\alpha})}(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}(1+\frac{\mu_i^{-c}}{\alpha})^{-y_{ij}}, & \text{for } y_{ij} > 0. \end{cases} \tag{4.3.8}$$

This model shares the same mean as the ZIP, but has variance $\text{Var}(y_{ij}|x_i)=(1-p_i)\mu_i(1+\alpha\mu_i^c+p_i\mu_i)$. As $\alpha \to 0$, it can be shown that the ZINB reduces to the ZIP. The variance suggests that the ZINB exhibits overdispersion when $\alpha > 0$ or $p_i > 0$. In contrast to (4.3.2), the ZINB dispersion is represented by

$$\frac{\text{Var}(y_{ij}|x_i)}{E(y_{ij}|x_i)} = \frac{(1-p_i)\mu_i(1+\alpha\mu_i^c+p_i\mu_i)}{(1-p_i)\mu_i} = 1+\alpha\mu_i^c+p_i\mu_i. \tag{4.3.9}$$

The ZINB log-likelihood can be expressed as

$$\ell(\mu_i, p_i, \alpha) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(I_{(y_{ij}=0)}ln\left[p_i + (1-p_i)(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}\right] + I_{(y_{ij}>0)}ln\left[(1-p_i)\frac{\Gamma(y_{ij}+\frac{\mu_i^{1-c}}{\alpha})}{y_{ij}!\Gamma(\frac{\mu_i^{1-c}}{\alpha})}(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}(1+\frac{\mu_i^{-c}}{\alpha})^{-y_{ij}}\right]\right).$$

It follows that the score equation for $p_i$ is then

$$\begin{aligned}\frac{\partial\ell}{\partial p_i} &= \sum_{j=1}^{n_i}\left(I_{(y_{ij}=0)}\left[\frac{1-(1+\alpha\mu_i^c)^{\frac{-\mu_i^{1-c}}{\alpha}}}{(1-p_i)(1+\alpha\mu_i^c)^{\frac{-\mu_i^{1-c}}{\alpha}}+p_i}\right] - I_{(y_{ij}>0)}\left[\frac{1}{1-p_i}\right]\right) \\ &= \frac{N_0\left(1-(1+\alpha\mu_i^c)^{\frac{-\mu_i^{1-c}}{\alpha}}\right)}{(1-p_i)(1+\alpha\mu_i^c)^{\frac{-\mu_i^{1-c}}{\alpha}}+p_i} - \frac{N-N_0}{1-p_i},\end{aligned}$$

yielding the MLE for $p_i$ as

$$\hat{p}_i = \frac{N(1+\hat{\alpha}\hat{\mu}_i^c)^{\frac{-\hat{\mu}_i^{1-c}}{\hat{\alpha}}} - N_0}{N\left((1+\hat{\alpha}\hat{\mu}_i^c)^{\frac{-\hat{\mu}_i^{1-c}}{\hat{\alpha}}}-1\right)}. \tag{4.3.10}$$

In the same manner as for the ZIP MLE of $\mu_i$, substituting (4.3.6) into (4.3.10) provides the equation:

$$\hat{\mu}_i(N - N_0) - S\left(1 - (1 + \hat{\alpha}\hat{\mu}_i^c)^{\frac{-\hat{\mu}_i^{1-c}}{\hat{\alpha}}}\right) = 0, \tag{4.3.11}$$

where an estimate for the overdispersion parameter, $\hat{\alpha}$, can be obtained through optimisation methods or extracted from the ZINB function. By direct comparison, we may notice similarities between (4.3.7) and (4.3.11). One can show through the limit of $(1 + \hat{\alpha}\hat{\mu}_i^c)^{\frac{-\hat{\mu}_i^{1-c}}{\hat{\alpha}}}$ that (4.3.11) converges to (4.3.7) as $\hat{\alpha} \to 0$. This can also be observed through the ZINB density in (4.3.8) which becomes equivalent to a ZIP for infinitesimal values of $\alpha$.

### 4.3.3 Purpose of the zero-inflation parameter for PBI

The zero-inflation parameter, $p_i$, will be modelled according to three different scenarios. Firstly, through a logistic regression but with the proportion of the mixture assumed to be constant:

$$\text{logit}(p_i) = \gamma_0, \tag{4.3.12}$$

secondly, $p_i$ modelled as a linear function of the dose

$$\text{logit}(p_i) = \gamma_1 x_i, \tag{4.3.13}$$

and finally as in (4.3.13) but with an intercept included:

$$\text{logit}(p_i) = \gamma_0 + \gamma_1 x_i. \tag{4.3.14}$$

The value of $\gamma_1$ depends on the type of radiation and its capacity to damage the cells whereas $\gamma_0$ is related to the fraction of irradiated blood. Estimates for $p_i$ and $\text{SE}(p_i)$ can be obtained through maximum likelihood estimation. In the case of a single patient dose sample, i.e. $\hat{p}_i \equiv \hat{p}$ modelled through (4.3.12), the proportion of irradiated scored cells (fraction of exposure), which we denote by $F$, can be estimated via

$$\hat{F} = 1 - \hat{p}, \tag{4.3.15}$$

where clearly $\text{SE}(\hat{F}) = \text{SE}(\hat{p})$. We note that (4.3.15) is a simplifying assumption as it ignores certain effects (such as cell death) which prevent irradiated cells being observable at the time of scoring. We will investigate this claim further in Section 4.4.

## Example: Application to whole-body H2AX calibration data

For the construction of calibration curves, and from a modelling perspective, the quantity of interest is always the aberration yield (or aberration/cell). Figures 2.1 and 2.2 depict examples of Poisson dose-response curves, fitted separately as in (2.2.3) and combined using (2.2.4), to PHE-Foci1 H2AX whole-body 4h and 24h calibration data. The resulting parameter estimates and their associated QP standard errors are reported below in Table 4.1. The computation of the standard errors in the combined models is shown in Appendix A.1. This procedure leads to smaller parameter standard errors as compared to the separate models. The dispersion estimates indicate present overdispersion for both timepoints (much greater at 24h post-exposure) but not mass overdispersion ($\hat{\phi} < 2$). The square root of the dispersion magnitudes are reflected in the increased quasi-Poisson parameter uncertainties, as compared with their Poisson equivalents.

| Fit type | Time | $\hat{\boldsymbol{\beta}} \pm \mathrm{SE}_{QP}(\hat{\boldsymbol{\beta}})$ | $\hat{\phi}$ |
|---|---|---|---|
| Linear (sep) | 4h | $(0.766 \pm 0.042,\ 1.700 \pm 0.058)$ | 1.444 |
| | 24h | $(0.333 \pm 0.032,\ 1.034 \pm 0.050)$ | 1.932 |
| Quadratic (sep) | 4h | $(0.700 \pm 0.040,\ 2.703 \pm 0.175,\ -0.412 \pm 0.065)$ | 1.411 |
| | 24h | $(0.311 \pm 0.031,\ 1.623 \pm 0.151,\ -0.251 \pm 0.057)$ | 1.915 |
| Linear (comb) | 4h | $(0.541 \pm 0.027,\ 1.840 \pm 0.061)$ | 1.693 |
| | 24h | $(0.541 \pm 0.027,\ 0.920 \pm 0.046)$ | |
| Quadratic (comb) | 4h | $(0.499 \pm 0.026,\ 3.017 \pm 0.185,\ -0.496 \pm 0.070)$ | 1.665 |
| | 24h | $(0.499 \pm 0.026,\ 1.400 \pm 0.141,\ -0.196 \pm 0.053)$ | |

Table 4.1 Fitted models to the PHE-Foci1 full-exposure data showing fit type, timepoint, Poisson/quasi-Poisson coefficient values and the corresponding standard errors and dispersion values for the quasi-Poisson regression. We note that standard errors are presented as opposed to $t$ statistics to allow comparison with alternative estimators in later tables.

It is well-reported that the relationship between experimental dose and focus counts is assumed to be linear, since increasing the dose linearly increases the number of electron tracks and ionisations that produce double–strand breaks [52]. However, for larger doses, H2AX foci have an increasing propensity to overlap, leading to a saturation effect [74]. It is clear from the negative parameter estimates of the quadratic terms that this is indeed the case, particularly at the 4h timepoint where foci yields per unit dose are higher. While there remains evidence for significance (all p-values $< 10^{-4}$), the variance contribution made by these quadratic terms can lead to imprecise

uncertainty of dose estimates [96]. For this reason, further analysis will be based on linear calibration curves only.

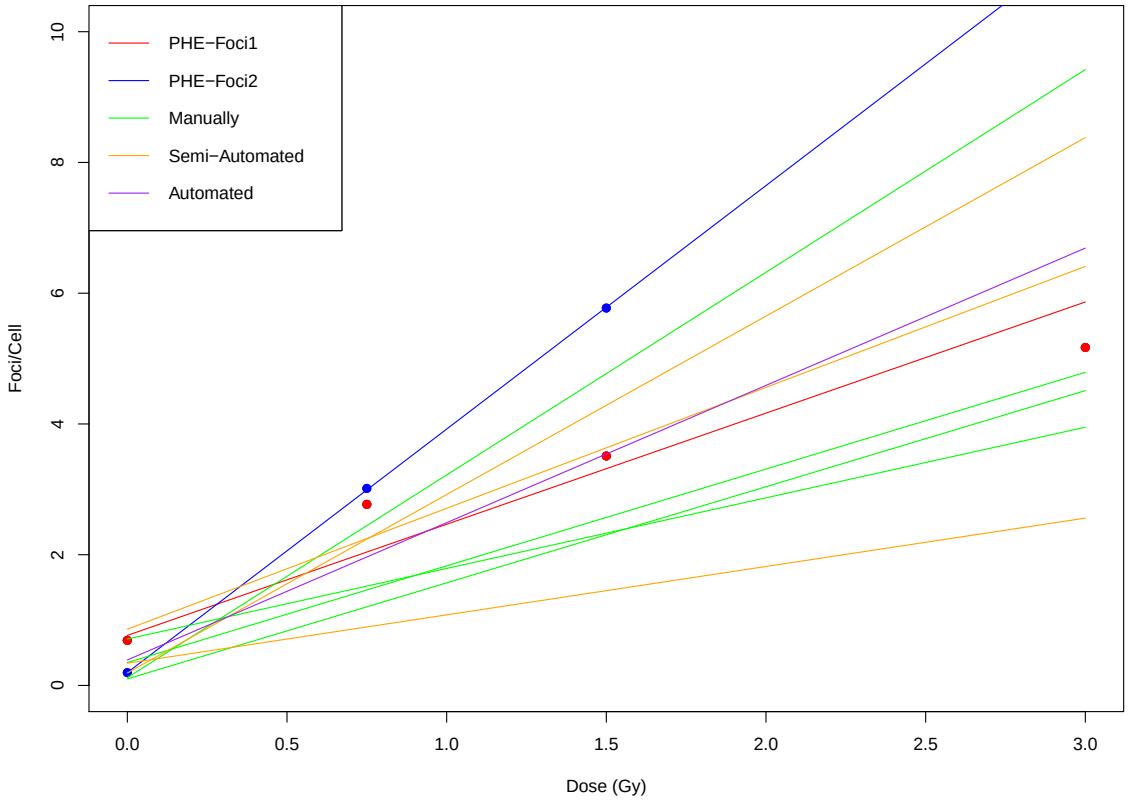



Fig. 4.1 A comparison of 4h calibration curves reported from various laboratories. Average $\gamma$-H2AX foci per cell as a function of 250kVp X-ray (red and blue lines) and Co-60 gamma-ray (green, orange and purple lines).

A comparison between our 4h dose–response curves, based on X-irradiation induced foci, with those from similar published studies that used $\gamma$-irradiation [94] revealed a sizable range of foci yields (as shown in Figure 4.1). Firstly, it can be seen that there is a significant difference in the estimated background level of foci between the PHE-Foci1 (refer to intercept term in first row of Table 4.1) and PHE-Foci2 ($\lambda_i = 0.197 \pm 0.014 + (3.724 \pm 0.029)x_i$) dataset. Additionally, this difference carries over to the dispersions, noting an estimated $\hat{\phi} = 1.079$ for the PHE-Foci2 calibration data (in contrast to $\hat{\phi} = 1.444$ as reported for the PHE-Foci1 data). The average background foci per cell for the Co-60 manually scored curves was found to be 0.32, suggesting the PHE-Foci1 background is perhaps relatively large. Secondly, and including the PHE datasets, foci yields were found to be largest on average for manual scoring, closely followed by automated then semi-automated. These findings are consistent with [103].

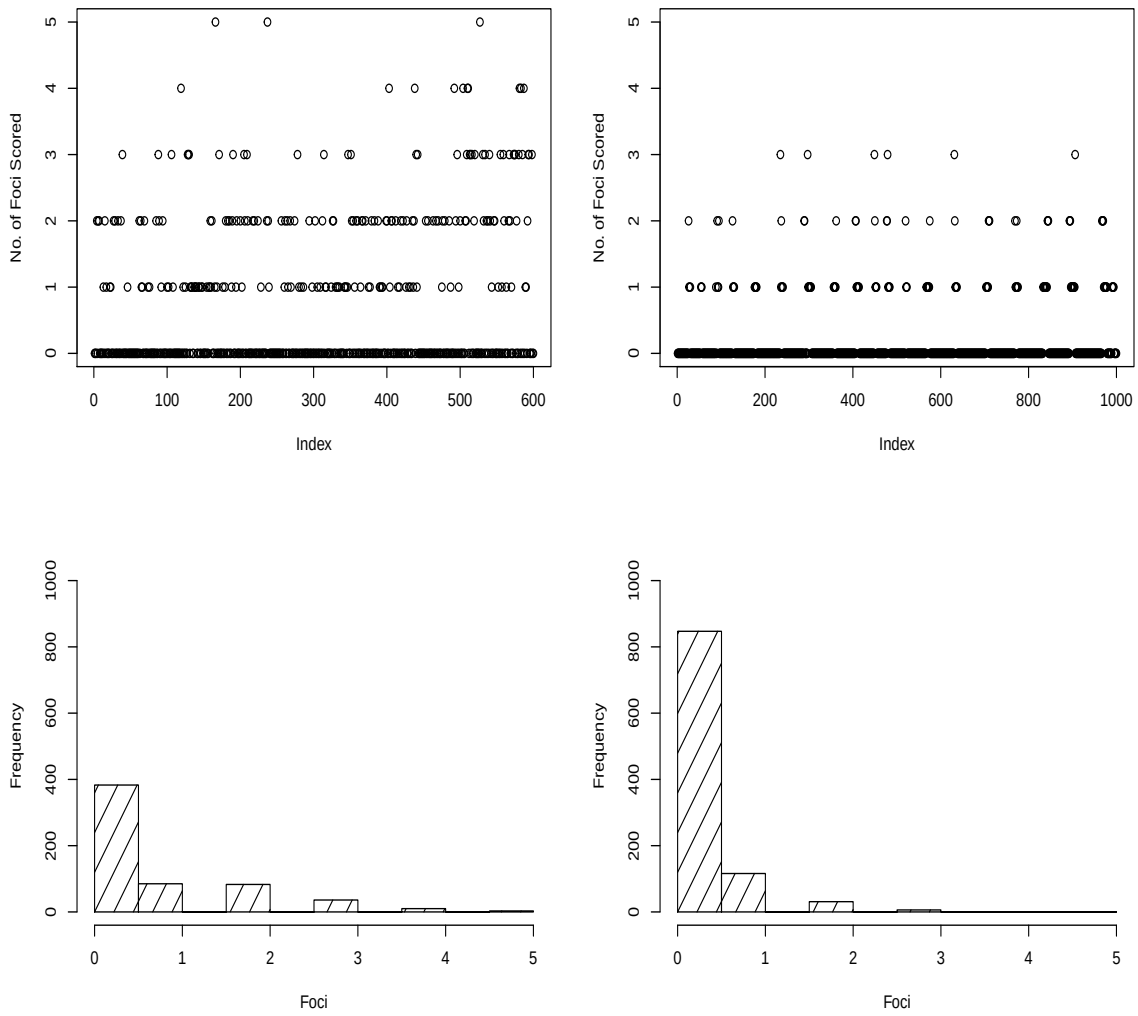



Fig. 4.2 Pair of plots with equal scales illustrating the foci counts (each count represented here by an index number) (top) and in the form of histograms (bottom) recorded for the PHE-Foci1 (left panels) and PHE-Foci2 (right panels) 4h 0Gy dose samples.

In a similar manner to (4.1.2), dispersion can also be computed using a deviance based estimation, that is

$$\hat{\phi}_{Dev} = \frac{\text{residual deviance}}{\nu}.$$

For the PHE-Foci1 and PHE-Foci2 datasets, these turn out to be $\hat{\phi}_{Dev} = 1.624$ and 1.029 respectively. As a rule of thumb, since $\text{SD}(\chi^2_\nu/\nu) = \sqrt{2/\nu}$, we say there is clear evidence of overdispersion if $\hat{\phi}_{Dev}$ exceeds 2 standard deviations of 1. This appears only to be true for the PHE-Foci1 dispersion, while the PHE-Foci2 dispersion remains within 1 SD ($1 + \sqrt{2/\nu} = 1.041$). We may wish to learn more information on the variability associated with the QP dispersion. Substituting our estimates for $\hat{\phi}$ into (4.1.6), we find that $\text{Var}(\hat{\phi}) = \dfrac{2 \times 1.444^2}{1200 - 2} = 0.003$ and $\text{Var}(\hat{\phi}) = \dfrac{2 \times 1.079^2}{4000 - 2} = 0.001$ providing $\hat{\phi} \in (1.328, 1.560)$ and $\hat{\phi} \in (1.031, 1.126)$. We note that the above computed $\hat{\phi}_{Dev}$ lie just outside these intervals suggesting slight disagreement between the QP and deviance-based dispersions.

Inter-laboratory differences can emanate from multiple sources such as the type and intensity of irradiation energy used (further research required), range of doses, technical or methodical variances, and scoring criteria, and the professional experience of the scorers. From initial information, we were made aware that a different scorer was responsible for the results from the latest PHE-Foci2 dataset. Previously noting the difference in foci backgrounds, in Figure 4.2 we plot separately the measured foci counts in the non-irradiated samples. Upon inspection, it appears that there exists some variation between the two scorers. The second scorer's maximum count is 3 foci, recorded just 6 times from 1000 cells, while the first scorer records a count greater than 3 much more than 6 times - they also have a maximum recorded count of 5 foci (based on the analysis of $3 \times 200$ cells 0Gy samples). For reference we will refer to the two scorers as "scorer $A$" and "scorer $B$" respectively.

To determine whether or not the variation between the two scorers is significant, we firstly carry out a Welch's two sample t-test for each level of dose. The test statistic remains

$$t_W = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

,

where our null hypothesis is that the population means of $A$ and $B$ are equal. From Table 4.2 all the $|t_W|$ correspond to p-values $< 0.05$, except slight evidence to retain the null hypothesis for the 0.75Gy samples. In addition, we also use the Mann-Whitney/Wilcoxen test statistic $t_M$ to investigate the difference in population medians denoted by $\tilde{y}_A$ and $\tilde{y}_B$. Again, it appears that only the 0.75Gy samples share a similar mean and median. We note that for the 0Gy samples, although it can be seen from Fig. 4.2 that both distributions are positively skewed, they behave differently for foci

| | Dose (Gy) | $\lvert\bar{y}_A - \bar{y}_B\rvert$ | $\lvert\tilde{y}_A - \tilde{y}_B\rvert$ | $\lvert t_W\rvert$ | Welch 95% CI | $\lvert t_M\rvert$ |
|---|---|---|---|---|---|---|
| | 0 | 0.494 | 0 | 10.59 ($2.2 \times 10^{-16}$) | (0.402, 0.586) | 230767 ($2.2 \times 10^{-16}$) |
| | 0.75 | 0.244 | 0 | 1.856 (0.065) | (-0.015, 0.503) | 105747 (0.192) |
| | 1.5 | 2.263 | 2 | 12.63 ($2.2 \times 10^{-16}$) | (1.911, 2.615) | 147803 ($2.2 \times 10^{-16}$) |
| | 3 | 6.19 | 6 | 33.32 ($2.2 \times 10^{-16}$) | (5.825, 6.555) | 191603 ($2.2 \times 10^{-16}$) |
| Overall | | 2.832 | 2 | 28.08 ($2.2 \times 10^{-16}$) | (2.635, 3.030) | 3221705 ($2.2 \times 10^{-16}$) |

Table 4.2 Welch and Mann-Whitney/Wilcoxen test statistic values and 95% confidence intervals (CI) for comparison of the individual dose samples and complete data. Associated p-values are given in parenthesis.

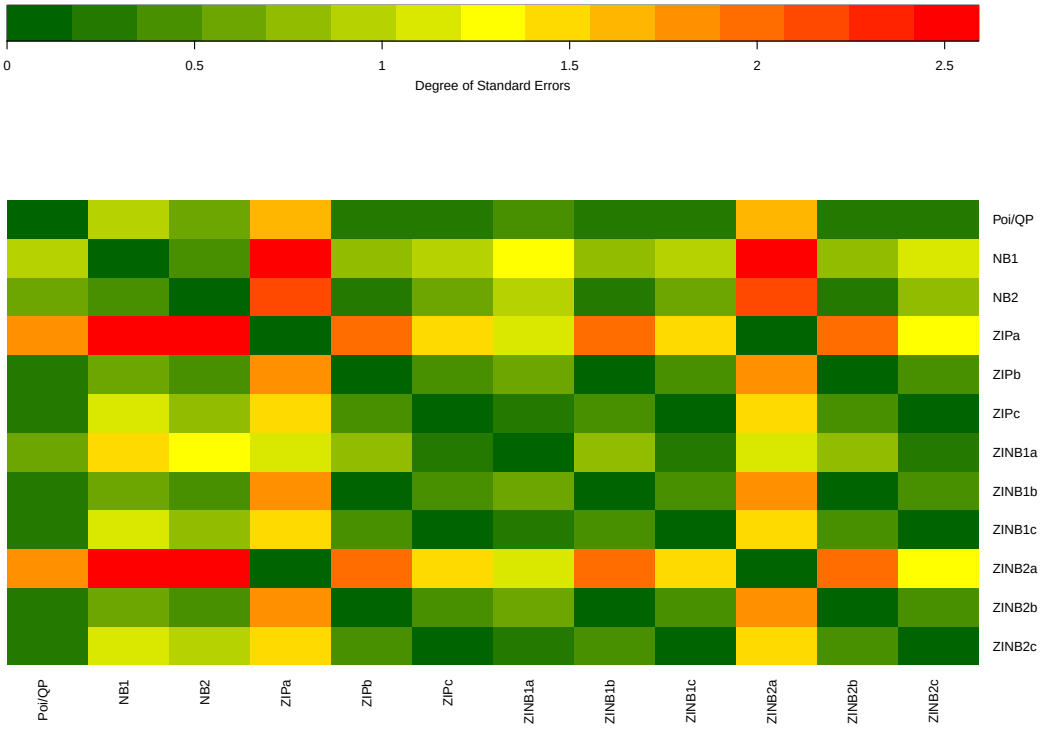
counts larger than 0 (most evident in comparing cells consisting of 1 and 2 foci) and it is preferable in such cases to summarise any difference based on the means rather than medians.

We are now interested to see how the estimates obtained from the Poisson regression and quasi-Poisson differ when we employ the NB models and their zero-inflated counterparts. Estimates for these models are shown in Table 4.3 under their individual assumptions about the variance of $y_{ij}$, where it is assumed that the conditional mean is correctly specified as in (2.2.2). Results obtained under ZIPa/ZINB1a/ZINB2a, ZIPb/ZINB1b/ZINB2b and ZIPc/ZINB1c/ZINB2c are based on the zero-inflation parameter being modelled as in (4.3.12), (4.3.13) and (4.3.14) respectively.

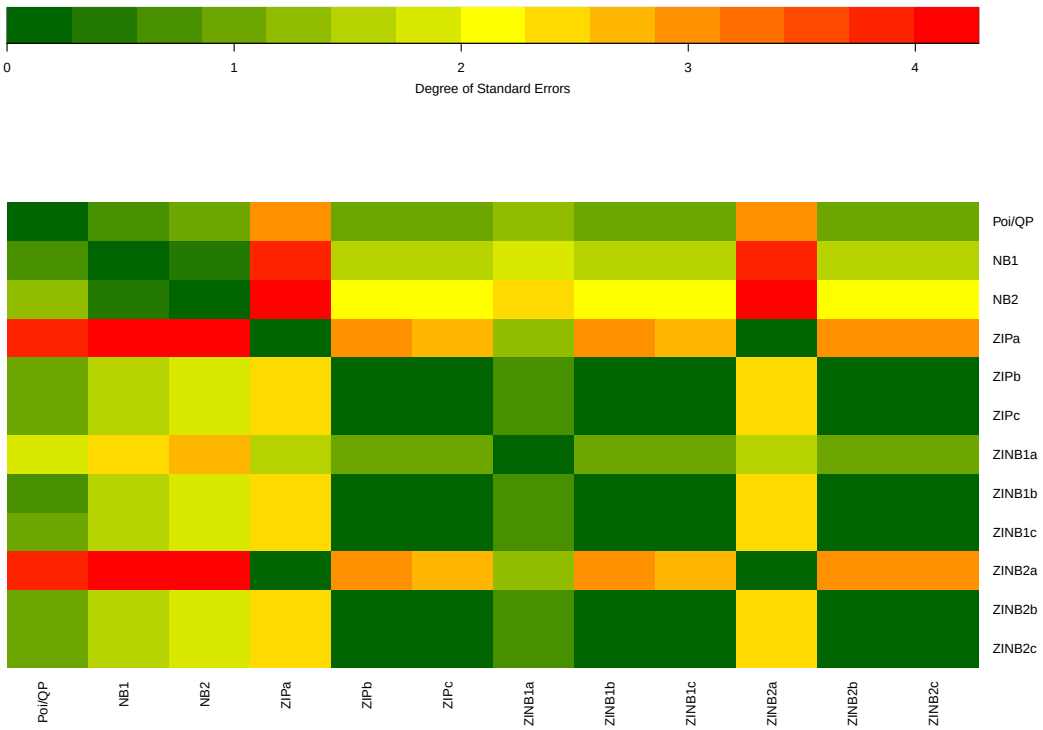
Comparing the previously estimated calibration curves and their standard errors from the quasi-Poisson with the various estimators in Table 4.3, it appears that they are mostly in agreement. The sensitivity analysis in Figures 4.3 and 4.4 (note: the base model is displayed on the horizontal axis while the compared model is on the vertical axis, for example we found for the PHE-Foci1 dataset that $\hat{\beta}_0 = 0.728$ from NB1 model is within 2.53 SEs of $\hat{\beta}_0 = 0.842$ from ZIPa model - as highlighted in solid red in Figure 4.3(a) - note the differing scales used for the degree of SEs between Figs 4.3(a), 4.3(b), 4.4(a), 4.4(b)) is within given showcases some deviations in coefficient values, mainly the ZIPa and ZINB2a models for the PHE-Foci1 data and the ZINB2b being the least comparable for the PHE-Foci2 data. In both datasets, the ZIPa model provided the largest intercept and smallest slope terms. To 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. For the NB1, we find that the value of $\hat{\alpha}$ in PHE-Foci1 is almost 10 times the reported estimate in PHE-Foci2. This

| | Model | $\hat{\boldsymbol{\beta}} \pm \text{SE}(\hat{\boldsymbol{\beta}})$ | $\hat{\alpha} \pm \text{SE}(\hat{\alpha})$ | $\hat{\gamma}_0$ | $\hat{\gamma}_1$ |
|---|---|---|---|---|---|
| PHE-Foci1 | NB1 | $(0.728 \pm 0.044,\ 1.743 \pm 0.062)$ | $0.619 \pm 0.084$ | | |
| | NB2 | $(0.744 \pm 0.038,\ 1.772 \pm 0.068)$ | $0.173 \pm 0.033$ | | |
| | ZIPa | $(0.842 \pm 0.045,\ 1.481 \pm 0.070)$ | | -1.533 | |
| | ZIPb | $(0.758 \pm 0.042,\ 1.650 \pm 0.057)$ | | | -1.021 |
| | ZIPc | $(0.775 \pm 0.048,\ 1.644 \pm 0.058)$ | | -0.086 | -0.973 |
| | ZINB1a | $(0.790 \pm 0.048,\ 1.589 \pm 0.080)$ | $0.153 \pm 0.182$ | -1.785 | |
| | ZINB1b | $(0.758 \pm 0.043,\ 1.651 \pm 0.057)$ | $< 0.001$ | | -1.021 |
| | ZINB1c | $(0.775 \pm 0.048,\ 1.645 \pm 0.058)$ | $< 0.001$ | -0.086 | -0.973 |
| | ZINB2a | $(0.842 \pm 0.045,\ 1.481 \pm 0.070)$ | $< 0.001$ | -1.534 | |
| | ZINB2b | $(0.758 \pm 0.043,\ 1.650 \pm 0.057)$ | $< 0.001$ | | -1.024 |
| | ZINB2c | $(0.779 \pm 0.047,\ 1.647 \pm 0.058)$ | $< 0.001$ | -0.187 | -0.957 |
| PHE-Foci2 | NB1 | $(0.191 \pm 0.014,\ 3.729 \pm 0.027)$ | $0.064 \pm 0.022$ | | |
| | NB2 | $(0.197 \pm 0,\ 3.725 \pm 0)$ | $< 0.001$ | | |
| | ZIPa | $(0.199 \pm 0.014,\ 3.712 \pm 0.029)$ | | -4.093 | |
| | ZIPb | $(0.193 \pm 0.015,\ 3.729 \pm 0.029)$ | | | -3.086 |
| | ZIPc | $(0.197 \pm 0.014,\ 3.725 \pm 0.029)$ | | -2.724 | -0.724 |
| | ZINB1a | $(0.198 \pm 0.014,\ 3.713 \pm 0.030)$ | $0.006 \pm 0.247$ | -4.101 | |
| | ZINB1b | $(0.193 \pm 0.015,\ 3.728 \pm 0.029)$ | $< 0.001 \pm 0.084$ | | -3.061 |
| | ZINB1c | $(0.197 \pm 0.014,\ 3.725 \pm 0.029)$ | $< 0.001 \pm 0.200$ | -2.667 | -0.758 |
| | ZINB2a | $(0.191 \pm 0.013,\ 3.741 \pm 0.030)$ | $< 0.001$ | -4.161 | |
| | ZINB2b | $(0.167 \pm 0.013,\ 3.790 \pm 0.029)$ | $< 0.001$ | | -2.412 |
| | ZINB2c | $(0.188 \pm 0.013,\ 3.745 \pm 0.029)$ | $< 0.001$ | -1.658 | -1.775 |

Table 4.3 Results of fitting various models to 4h post-exposure whole-body calibration data for datasets PHE-Foci1 and PHE-Foci2.

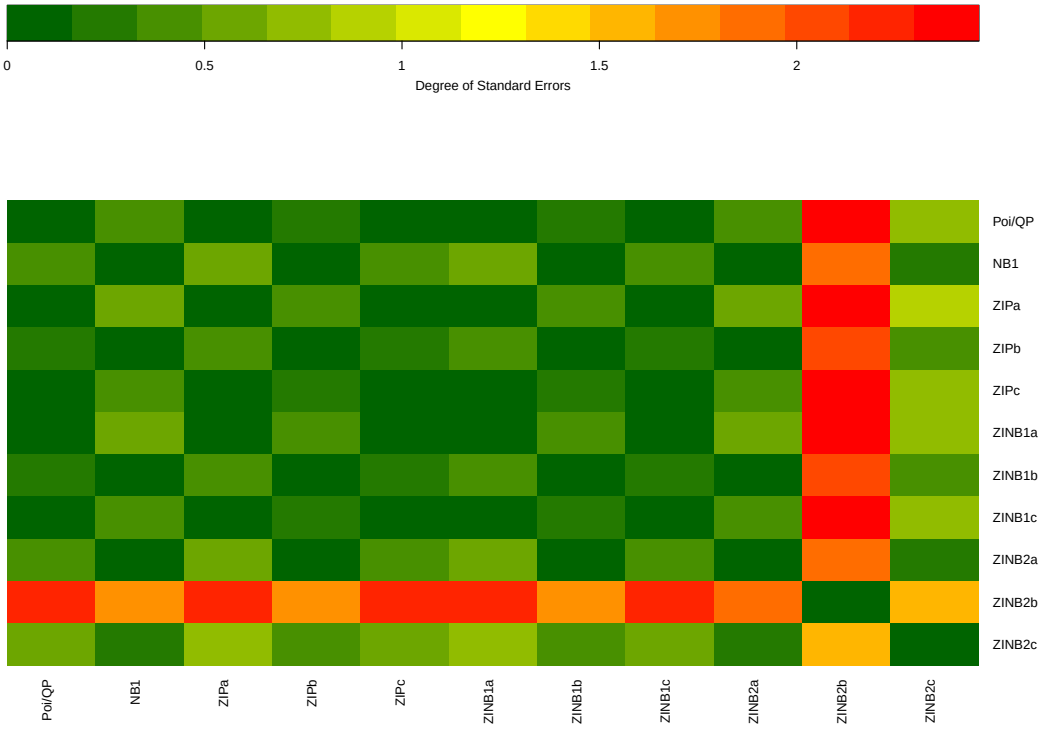



(a) Accuracy of $\hat{\beta}_0$ as compared to each model.

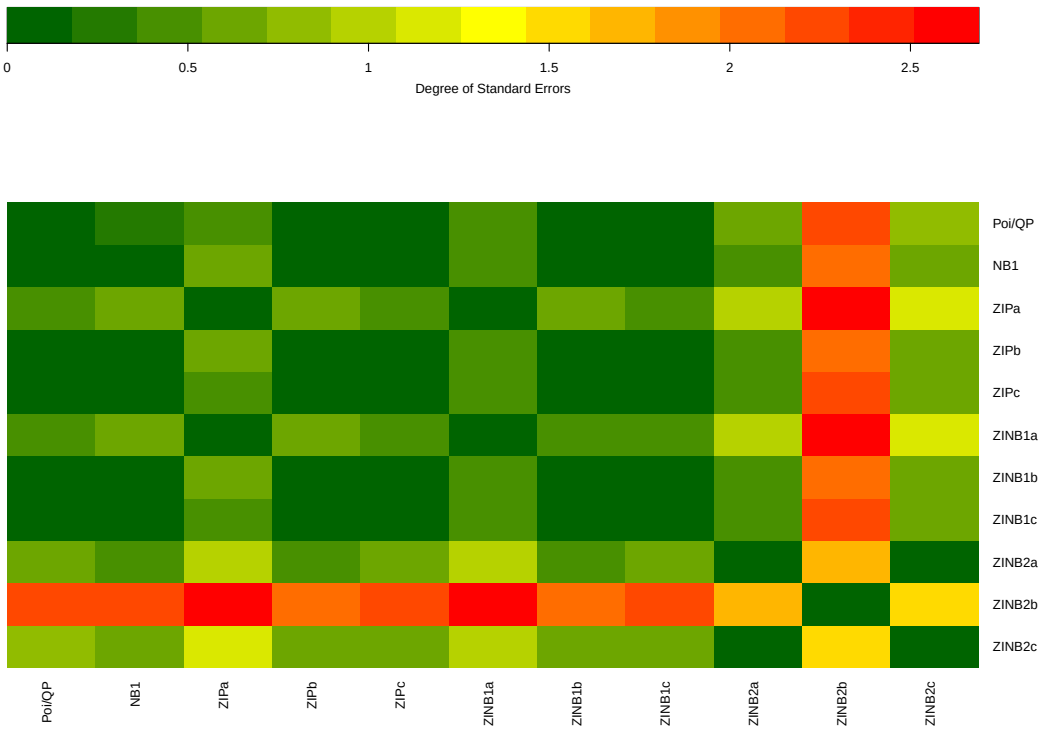



(b) Accuracy of $\hat{\beta}_1$ against each model (PHE-Foci1).

Fig. 4.3




(a) Accuracy of $\hat{\beta}_0$ as compared to each model.




(b) Accuracy of $\hat{\beta}_1$ against each model (PHE-Foci2).

Fig. 4.4

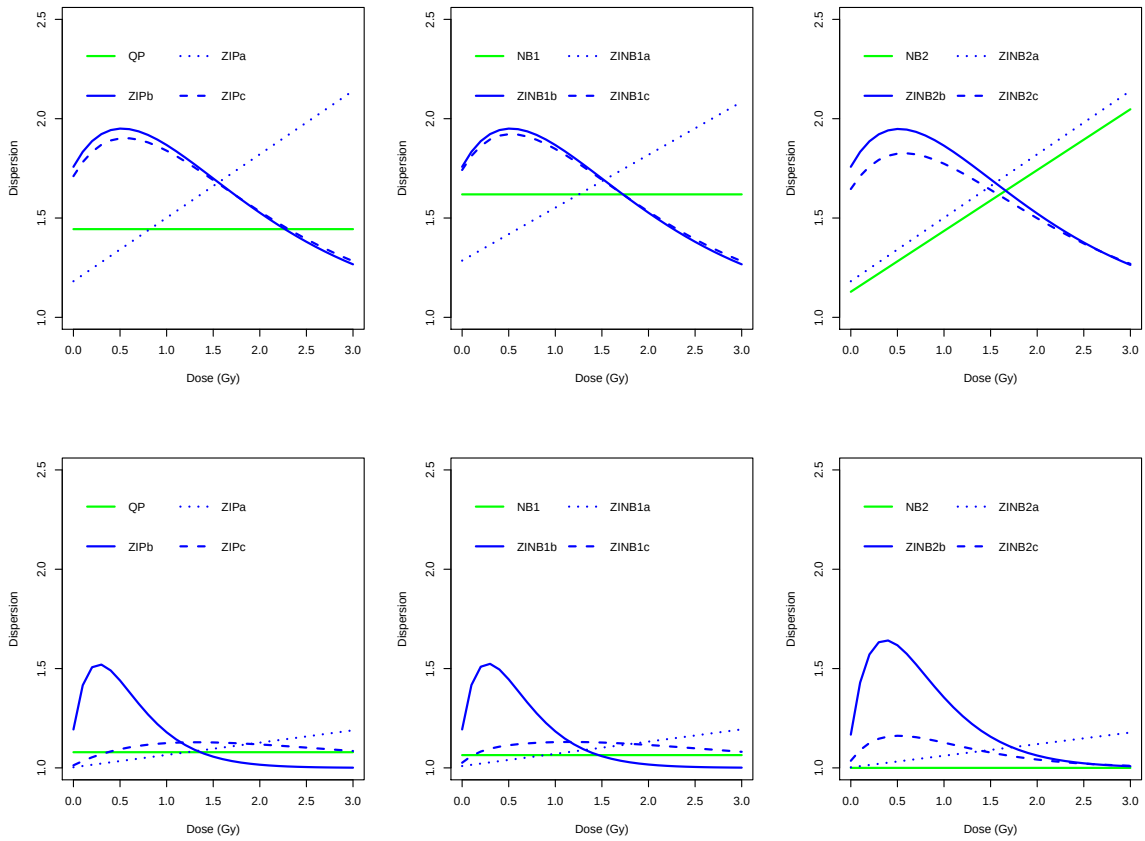



Fig. 4.5 Dose vs dispersion behaviour for PHE-Foci1 (top) and PHE-Foci2 (bottom) comparing QP with ZIP (1st column), NB1 with ZINB1 (2nd column) and NB2 with ZINB2 (final column).

can be viewed as a consequence of the underdispersion in the 3Gy sample which we observed in Chapter 3.

It is particularly convenient to compare the dispersion from an NB1 model with $\hat{\phi}$ acquired from a quasi-Poisson. Interestingly, the PHE-Foci1 NB1 dispersion is greater ($1.619 > 1.444$) but is smaller for PHE-Foci2 ($1.064 < 1.079$). Comparing further the $\hat{\alpha}$ from the ZINB1a models, a value of 0.153 indicates that there must be some unobserved heterogeneity, in addition to zero-inflation, contributing to the PHE-Foci1 dispersion. The PHE-Foci2 estimate of 0.006 suggests the majority of PHE-Foci2 dispersion is attributed to zero-inflation, as somewhat explained by the more negative $\hat{\gamma}_0$. We note here that since $\hat{\alpha}$ operates on the boundary of the parameter space, the corresponding standard errors are considered unreliable and are merely stated for completeness, where possible. It is evident in both datasets that when the zero-inflation is modelled with the $\gamma_1$ constant (i.e. with dependence on covariate dose), the dispersion is fully explained by $p_i$, rendering $\alpha$ insignificant.

Due to their covariate-dependent dispersions, the remaining models presented in Table 4.3 are not easily comparable with the NB1 and quasi-Poisson model. Figure

| Dose (Gy) | | NB2 | ZIPa | ZIPb | ZIPc | ZINB1a | ZINB1b | ZINB1c | ZINB2a | ZINB2b | ZINB2c |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.75 | QP | -5.9% | **-1.6%** | 33.5% | 30.8% | 2.9% | 33.5% | 31.7% | **-1.6%** | 33.3% | 25.7% |
| | NB1 | -16.1% | -12.2% | 19.1% | 16.6% | **-8.3%** | 19.1% | 17.5% | -12.2% | 18.9% | 12.1% |
| 1.5 | QP | **10.0%** | 15.0% | 17.7% | 17.1% | 16.7% | 17.7% | 17.3% | 15.0% | 17.4% | 13.7% |
| | NB1 | -1.9% | 2.6% | 5.0% | 4.5% | 4.1% | 4.9% | 4.6% | 2.6% | 4.7% | **1.4%** |
| 3 | QP | 41.8% | 48.2% | -12.3% | **11.2%** | 44.4% | -12.3% | **-11.2%** | 48.2% | -12.5% | -12.1% |
| | NB1 | 26.5% | 32.2% | -21.8% | **-20.8%** | 28.8% | -21.8% | **-20.8%** | 32.2% | -21.9% | -21.6% |

Table 4.4 % difference between constant and non-constant dispersions evaluated at dose levels considered (PHE-Foci1).

| Dose (Gy) | | NB2 | ZIPa | ZIPb | ZIPc | ZINB1a | ZINB1b | ZINB1c | ZINB2a | ZINB2b | ZINB2c |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.75 | QP | -7.3% | -2.7% | 20.1% | 3.2% | **-2.2%** | 20.6% | 4.3% | -3.0% | 38.4% | 6.7% |
| | NB1 | -6.0% | -1.4% | 21.7% | 4.7% | **-0.8%** | 22.2% | 5.7% | -1.7% | 40.3% | 8.1% |
| 1.5 | QP | -7.3% | 1.6% | -2.1% | 4.6% | 2.1% | -1.9% | 4.5% | 1.1% | 7.2% | **-0.2%** |
| | NB1 | -6.0% | 3.0% | -0.7% | 6.0% | 3.5% | **-0.5%** | 5.9% | 2.4% | 8.7% | 1.2% |
| 3 | QP | -7.3% | 10.2% | -7.2% | 0.6% | 10.6% | -7.2% | **0.2%** | 9.2% | -6.6% | -6.3% |
| | NB1 | -6.0% | 11.7% | -5.9% | **1.9%** | 12.2% | -5.9% | 1.6% | 10.7% | -5.3% | -5.1% |

Table 4.5 % difference between constant and non-constant dispersions evaluated at dose levels considered (PHE-Foci2).

4.5 attempts to show the dispersion behaviour of the QP, NB1 and NB2 as compared to their zero-inflated counterpart models. Due to the small $\hat{\gamma}_0$, we notice that those models based on (4.3.13) and (4.3.14) produce similar dispersions for the PHE-Foci1 data (also noting similarities between the NB2 and ZINB2a). On the other hand, it is clear that the PHE-Foci2 dataset provides more distinct dispersion patterns. One could possibly argue that the behaviour of the model dispersions under (4.3.12) and (4.3.14) can be roughly captured by the QP and NB1 dispersions.

Let us assume that a laboratory provides a Poisson/QP or NB1-fitted calibration curve along with their dispersions, then it is desirable to see how they compare if a different model had been employed. To compliment Figure 4.5, Tables 4.4 and 4.5 can be used to assess the constant dispersion against dose-dependent dispersion estimated at each level of dose. It is evident from the values of $\hat{\delta}$ in Chapter 3 that dispersion increases with dose in PBI, however this relationship does not occur for WBI. At 0.75Gy, we see for the PHE-Foci1 that the QP dispersion almost replicates that of the ZIPa and ZINB2a while the NB1 dispersion is most close to the ZINB1a. Interestingly, for both datasets the highest dispersion was recorded for the 1.5Gy samples, however all panels in Figure 4.5 reveal that the dispersion either peaks early (around a dose of 0.5Gy) or continues to increase with dose. The ZIPc and ZINB1c dispersions in the PHE-Foci2 are the only exceptions, both reaching a maximum at approx 1.5Gy.

It can be observed from the estimates of $\gamma_0$ and $\gamma_1$ that the behaviour of the zero-inflation parameter $p_i$ can vary depending on the specified model. Figure 4.6 shows the fitted values of $p_i$ after fitting ZIP, ZINB1 and ZINB2 regression models to the aforementioned whole-body calibration data and the partially-exposed calibration data (for the exposure levels considered in Tables 3.1 and 3.2). The solid dots represent the fitted $p_i$ when these do not depend on covariates, and the dashed and solid lines signify $p_i$ modelled through a logit link as a linear function of the dose with and without intercept respectively. It is clear the value of $p_i$ is influenced by the percentage of unirradiated blood. Moreover, in all models, $p_i$ takes very similar values for larger doses. However, this is not the case for the lowest doses, particularly for the PHE-Foci2 data. The dashed lines in top panels suggest that for non–irradiated blood samples, the probability of extra zeros is quite similar. However, the dashed lines in the bottom plots show that the probability takes very different values at dose 0Gy. We discover that only in the case of 30% exposure for PHE-Foci2 that a positive $\hat{\gamma}_1$ is obtained, hence $p_i$ increases (as opposed to decreases) with dose.

When we identify the possible presence of overdispersion, what are the consequences of failing to take it into account? We observed that the standard errors obtained from a Poisson model will be incorrect and may be seriously underestimated and consequently we may incorrectly assess the significance of individual regression parameters. To overcome this, it is preferable to utilise either a quasi-Poisson or a NB model. Often choosing between the quasi-Poisson and NB1 is attributed to preference, however

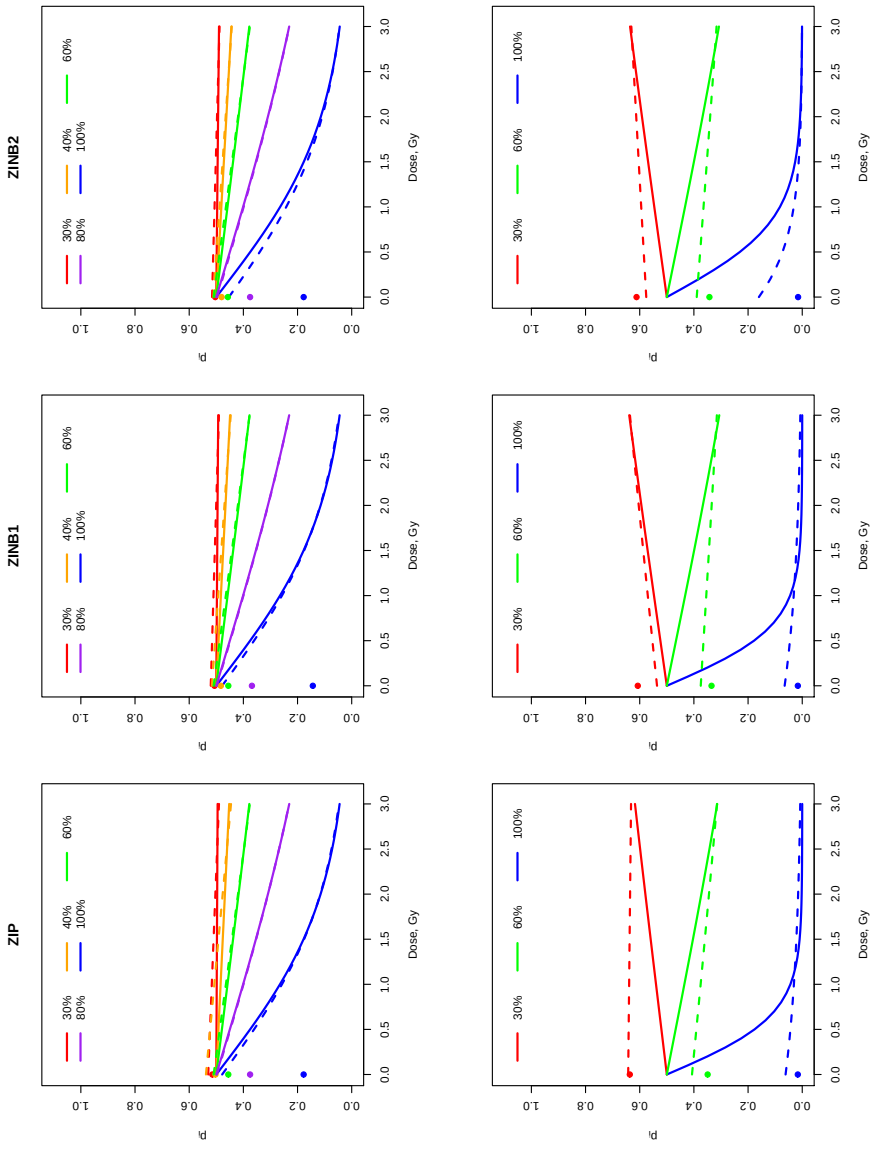



Fig. 4.6 Fitted zero–inflation (mixture) parameters $p_i$ as a function of dose, $x_i$, to full- and partial-exposure calibration data for PHE-Foci1 (top panels) and PHE-Foci2 (bottom panels). Solid lines correspond to modelling the mixture parameter as $\text{logit}(p_i) = \gamma_1 x_i$ and dashed lines correspond to $\text{logit}(p_i) = \gamma_0 + \gamma_1 x_i$. Solid dots indicate the fitted probabilities when $\text{logit}(p_i) = \gamma_0$.

differences can arise in their dispersions. A distinct advantage of the NB is that the maximum likelihood approach allows the standard likelihood tests to be implemented. In Chapter 5 we will introduce tests for overdispersion, specifically we discuss how the overdispersion parameter, $\alpha$, can be used as information against a Poisson fit.

## 4.4 Impact of cell death

Early research has suggested that radiation overexposure accidents where the doses are up to 0.5Gy, the question of differences in transformation and survival between non-irradiated cells and those irradiated is not a complicating factor. However, for higher doses it has been reported that some allowance for cell death should be considered when acute partial body irradiation is known to have occurred [67]. For the $\gamma$-H2AX biomarker, cells are usually scored after a few hours, which leaves much less time for cell death than for the dicentric biomarker, where at least 48 hours need to pass until mitosis [47]. While it appears, on this basis, reasonable to assume that for the $\gamma$-H2AX biomarker the original irradiated fraction corresponds to the fraction of irradiated cells at the time of scoring, we still would like to investigate this claim further.

We recall from [47] that the corrected fraction of irradiated cells can be written as

$$F = \frac{1-p}{1-p+p\kappa} \tag{4.4.1}$$

where $\kappa = \kappa(D)$ is a dose-dependent function describing the survival rate of irradiated cells. According to Lloyd and Edwards [66], this rate follows a decreasing exponential function of the dose $x$,

$$\kappa(x) = e^{-\gamma_1 x}. \tag{4.4.2}$$

In Hilali et al's context [47] of dicentric chromosomes, they denote $\gamma_1 = 1/x_0$, where $x_0$ can be interpreted as the initial dose required to reduce the number of irradiated cells to 37% due to interphase death or mitotic delay. The range of plausible values for $x_0$ for this biomarker has been postulated in the literature, without much justification, to be between 2.7 and 3.5Gy [67].

From (4.4.1) it is clear that we have $F = 1 - p$ exactly when $\kappa = 1$, i.e. when the survival rate is approximately 100%. From (4.4.2), this implies $\gamma_1$ approaching 0 (or $x_0$ tending to infinity). Oliveira et al. [76] demonstrated that when modelling the proportion $p$ via (4.3.14), then the constant $\gamma_1$ in (4.4.2) corresponds to $\gamma_1$ in (4.3.14). So, it remains to show that, $\gamma_1$ is not statistically different from 0. From the zeroes contained in the $\hat{\gamma}_1$ confidence intervals quoted in Table 4.6, it is clear that effects such as cell death at the time of scoring can be considered negligible and therefore provides sufficient evidence for the exposed fraction to be estimated through $\hat{F} = 1 - \hat{p}$

in the case of a single sample. The right-hand column of the table states the individual confidence intervals for $\kappa(x)$ for each dose. Substituting any of the three doses leads to confidence intervals for $\kappa(x)$ which encompass a value of 1.

Table 4.7 reveals the above conclusions made for the PHEFoci1 dataset hold only in the case of the ZIP and ZINB2 models fitted to the 30% exposure samples from the PHEFoci2 dataset. The remaining ZINB1 yields positive confidence limits for $\hat{\gamma}_1$ and hence $\kappa(x) < 1$, implying dose and fraction estimates will need to be updated using (4.4.1) to justify for some cell deterioration. From a biological perspective, it remains that $\gamma_1$ (and $x_0$) should be strictly positive. In reference to (4.4.2), the maximum survival rate is achieved when $\gamma_1 = 0$ and we therefore interpret the observed negative $\hat{\gamma}_1$ values for the 60% exposure as $\kappa(x) = 1$. Furthermore, the reported confidence intervals in Table 4.8 indicate here we do not need to worry about cell death in our dicentric data.

Table 4.6 99% confidence intervals for $\gamma_1$ and $\kappa(x)$ for PHE-Foci1 data (omitting control/0Gy sample).

| $F$ (%) | Model | $\hat{\gamma}_1$ | $\kappa(x)$ | | |
|---|---|---|---|---|---|
| | | | $x = 0.75$ | $x = 1.5$ | $x = 3$ |
| 30 | ZIP | (-0.148, 0.316) | (0.789, 1.117) | (0.622, 1.248) | (0.387, 1.557) |
| | ZINB1 | (-0.136, 0.347) | (0.771, 1.108) | (0.594, 1.227) | (0.353, 1.505) |
| | ZINB2 | (-0.142, 0.336) | (0.777, 1.112) | (0.604, 1.238) | (0.365, 1.533) |
| 40 | ZIP | (-0.092, 0.376) | (0.754, 1.071) | (0.569, 1.148) | (0.324, 1.318) |
| | ZINB1 | (-0.088, 0.400) | (0.741, 1.068) | (0.548, 1.141) | (0.301, 1.301) |
| | ZINB2 | (-0.089, 0.397) | (0.742, 1.069) | (0.551, 1.143) | (0.304, 1.307) |
| 60 | ZIP | (-0.104, 0.368) | (0.759, 1.081) | (0.576, 1.169) | (0.331, 1.368) |
| | ZINB1 | (-0.105, 0.367) | (0.760, 1.082) | (0.577, 1.172) | (0.332, 1.373) |
| | ZINB2 | (-0.105, 0.368) | (0.759, 1.081) | (0.576, 1.170) | (0.331, 1.368) |
| 80 | ZIP | (-0.192, 0.312) | (0.791, 1.155) | (0.627, 1.335) | (0.393, 1.782) |
| | ZINB1 | (-0.197, 0.307) | (0.794, 1.160) | (0.631, 1.345) | (0.398, 1.809) |
| | ZINB2 | (-0.199, 0.306) | (0.795, 1.161) | (0.632, 1.348) | (0.400, 1.816) |

Table 4.7 99% confidence intervals for $\gamma_1$ and $\kappa(x)$ for PHE-Foci2 data (omitting control/0Gy sample).

| $F$ (%) | Model | $\hat{\gamma}_1$ | $\kappa(x)$ | | |
|---|---|---|---|---|---|
| | | | $x = 0.75$ | $x = 1.5$ | $x = 3$ |
| 30 | ZIP | (-0.067, 0.119) | (0.915, 1.052) | (0.837, 1.106) | (0.700, 1.223) |
| | ZINB1 | (0.056, 0.300) | (0.798, 0.959) | (0.637, 0.919) | (0.406, 0.845) |
| | ZINB2 | (-0.009, 0.205) | (0.857, 1.007) | (0.735, 1.014) | (0.540, 1.028) |
| 60 | ZIP | (-0.275, -0.080) | | | |
| | ZINB1 | (-0.228, -0.020) | | | |
| | ZINB2 | (-0.246, -0.042) | | | |

Table 4.8 99% confidence intervals for $\gamma_1$ and $\kappa(x)$ for PHE-Dicentric data.

| $F$ (%) | Model | $\hat{\gamma}_1$ | $\kappa(x)$ | | |
|---|---|---|---|---|---|
| | | | $x = 0.5$ | $x = 0.7$ | $x = 1$ |
| 50 | ZIP | (-1.979, -0.324) | | | |
| | ZINB1 | (-2.197, 0.333) | (0.847, 2.999) | (0.792, 4.653) | (0.717, 8.995) |
| | ZINB2 | (-2.421, -0.179) | | | |
| 75 | ZIP | (-3.106, -1.046) | | | |
| | ZINB1 | (-3.534, -0.289) | | | |
| | ZINB2 | (-3.726, -0.825) | | | |

# Chapter 5

# Model-based overdispersion tests

The failure of the Poisson assumption of equidispersion has similar consequences to failure of the assumption of homoskedasticity in the linear regression model. However, the effect on reported standard errors and $t$ statistics can be much larger. Practical and reliable tests for overdispersion are important to justify the need for models beyond the standard Poisson regression. Various tests have been developed, and [49] provides a good review on this topic. In the Poisson regression framework, [64] developed a unifying theory and derived score tests for overdispersion with respect to both Poisson and binomial regression models, while the explicit forms of the test statistics are only given in certain special cases. The score statistics developed by [21] specifically for comparing the Poisson model against the negative binomial model, is a special case of the general score statistics later developed by [24]. Likelihood-based tests which make use of $\hat{p}_i$ and $\hat{\alpha}$, as previously estimated in Chapter 4, and model selection criteria are discussed and compared.

## 5.1 Test statistics for ZINB models

By noting that the ZINB distribution is a general model for counts which nests the ZIP, NB, and Poisson models, test statistics can be formed to detect either overdispersion, or zero-inflation or both simultaneously using the likelihood ratio test (LRT), Wald test, or score test. We will concentrate on the score test as this has the superior advantage of simpler calculation, in not requiring the model under the alternative hypothesis to be fitted and also leading to composite test statistics for the zero-inflation model. However, the other common test statistics will be summarised briefly.

### 5.1.1 Comparing ZIP and ZINB models

For testing a ZIP regression against ZINB alternatives the relevant hypothesis test is

$$H_0 : \alpha = 0 \qquad \text{against} \qquad H_1 : \alpha > 0.$$

The corresponding likelihood ratio test (LRT) statistic for detecting overdispersion in a ZIP model is given by:

$$R_{\alpha_c} = -2\left(\ell(\hat{\mu}_0, \hat{p}_0) - \ell(\hat{\mu}, \hat{\alpha}, \hat{p})\right) \tag{5.1.1}$$

where $\ell(\hat{\mu}_0, \hat{p}_0)$ and $\ell(\hat{\mu}, \hat{\alpha}, \hat{p})$ are the maximised log-likelihoods under the ZIP regression and the ZINB regression models, respectively. The associated Wald test statistic is

$$W_{\alpha_c} = \frac{\hat{\alpha}^2}{\widehat{\text{Var}}(\hat{\alpha})} \tag{5.1.2}$$

where $\widehat{\text{Var}}(\hat{\alpha})$ is the relevant diagonal element of the inverse ZINB Fisher information matrix, $I^{-1}(\hat{\mu}, \hat{\alpha}, \hat{p})$. The subscript $c = 0, 1$ here is used to identify the form of the NB model, for example $R_{\alpha_0}$ is the LRT for testing a ZIP regression against the ZINB1 model.

Under the null hypothesis, both $R_{\alpha_c}$ and $W_{\alpha_c}$ might be expected to have a $\chi^2_1$ distribution. Some care is required here as the null hypothesis is on the boundary of the parameter space (e.g. the null distribution of the LRT is not the usual $\chi^2(1)$ distribution), and also the alternative hypothesis is one-sided as we are only testing for overdispersion. With a notable exception being Lawless [64], this complexity is not usually discussed. A solution of hypothesis testing at boundary values was documented by Moran [75]. It is suggested that the asymptotic distribution associated with the LRT statistic has probability mass of one half at zero and a half $\chi^2(1)$ distribution above 0. If we are testing at level $\epsilon$, where $\epsilon > 0.5$, one rejects $H_0$ if the test statistic exceeds $\frac{1}{2}(\chi^2_{1-\epsilon}(0) + \chi^2_{1-\epsilon}(1))$ rather than $\chi^2_{1-\epsilon}(1)$. The Wald test is usually expressed as a $t$-test statistic, which is defined as having a mass of one half at zero and a normal distribution for values greater than zero. This means we continue to use the same critical value as for the LRT. For $\alpha$ operating on the boundary of the parameter space, the reference distributions of $R_{\alpha_c}$ and $W_{\alpha_c}$ are a mixture of a degenerate distribution at zero and a $\chi^2_1(\alpha > 0)$, with p-values for the LRT given by $\frac{1}{2}\text{P}(\chi^2_1 \geq R_{\alpha_c})$.

### 5.1.2 Comparing NB regression with ZINB models

Comparing NB models and ZINB regression corresponds to testing the hypotheses

$$H_0 : p_i = 0 \qquad \text{against} \qquad H_1 : p_i > 0.$$

For a general ZINB regression, the LRT for zero-inflation is

$$R_{p_{i_c}} = -2\left(\ell(\hat{\mu}_0, \hat{p}_0) - \ell(\hat{\mu}, \hat{\alpha}, \hat{p})\right) \tag{5.1.3}$$

where here $\ell(\hat{\mu}_0, \hat{p}_0)$ and $\ell(\hat{\mu}, \hat{\alpha}, \hat{p})$ are the maximised log-likelihoods under the NB and ZINB regression models. The associated Wald test statistic is

$$W_{p_{i_c}} = \hat{p}_i'\{\mathrm{Cov}(\hat{p}_i)\}^{-1}\hat{p}_i. \tag{5.1.4}$$

Under the null hypothesis, both $R_{p_{i_c}}$ and $W_{p_{i_c}}$ are asymptotically $\chi^2$ distributed. In the case of constant (5.1.4) reduces to the usual Wald test statistic

$$W_{p_{i_c}} = \frac{\hat{p}_i^2}{\widehat{\mathrm{Var}}(\hat{p}_i)}. \tag{5.1.5}$$

under the null hypothesis $p_i$ is on the boundary of the parameter space and again the appropriate reference distribution for $R_{p_{i_c}}$ and $W_{p_{i_c}}$ is an equal mixture of a constant at zero and a $\chi_1^2$ distribution, so for a test based on $W_{p_{i_c}}$ p-values are given by $\frac{1}{2}\mathrm{P}(\chi_1^2 \geq W_{p_{i_c}})$. For testing a Poisson against the ZIP, the variance of $p$ can be computed via

$$\mathrm{Var}(\hat{p}) = \frac{n_0\bar{y}(n_0 - n(\hat{\mu} - \bar{y})e^{-\hat{\mu}})}{n^2\hat{\mu}\left[(1 - e^{-\mu})(n_0 - n(\hat{\mu} - \bar{y})e^{-\hat{\mu}}) - n\hat{\mu}e^{-2\hat{\mu}}\right]},$$

of which the derivation is given in Appendix A.4.

## 5.2 Score tests for detecting zero-inflation

We learned from Chapter 3 that the present overdispersion in both datasets is mostly as a result of the excess zeros, suggesting a Poisson fit unsuitable. However, in order to investigate the degree of adequacy of the Poisson regression model in dealing with the incidence of zero counts, in Figure 5.1 we assess the probability of a zero, $\pi_0$, for various models with respect to the Poisson. For the NB1 and NB2 models the fitted zero probabilities are $\hat{\pi}_0(\hat{\mu}_i, \hat{\alpha}) = 1/(1 + \hat{\alpha})^{\frac{\hat{\mu}_i}{\hat{\alpha}}}$ and $\hat{\pi}_0(\hat{\mu}_i, \hat{\alpha}) = (1 + \hat{\alpha}\hat{\mu}_i)^{-\frac{1}{\hat{\alpha}}} \geq e^{-\hat{\mu}_i}$ respectively, where $\hat{\mu}_i$ (and $\hat{\alpha}$) have been estimated from that model.

To test whether an over-dispersed count distribution is zero-inflated, that is $H_0 : p_i = 0$, a score test can be used. Its test statistic defined as

$$\mathrm{Sc}(\theta_0) = \mathrm{Sc}(\theta_0)I(\theta_0)^{-1}\mathrm{Sc}(\theta_0)',$$

where Sc() is the score function, $I()$ is the Fisher information and $\theta_0$ is the MLE of the parameter set $\theta$ under $H_0$. In the considerations which follow we note that these tests

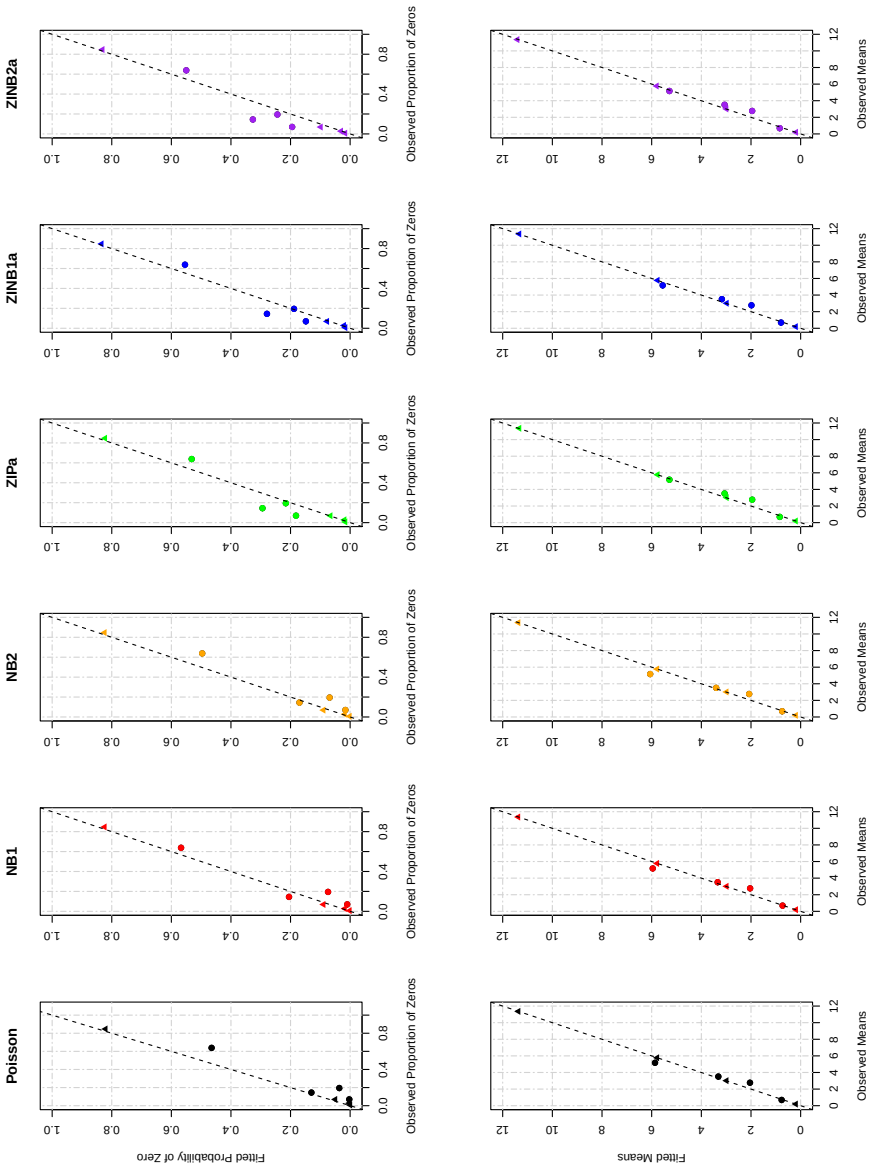



Fig. 5.1 Fitted vs observed proportion of zeros and means in each dose sample for the Poisson, NB and ZI ($p_i$ modelled as a constant) models. Circular points represent PHE-Foci1 4h samples and triangles the PHE-Foci2 0/100% exposure data. The dashed identity line is the Poisson base. Under the Poisson, the probability of a zero is inversely proportional to the mean, hence it is plausible for a large fraction of zeros with a small mean to maintain compatibility under a Poisson.

assume that $p_i \equiv p$ across observations. Furthermore, all tests require that the mean is modelled through a log-link function.

In 1995, van den Broek [102] presented a score test to determine whether the number of zeros in a sample is too large for a Poisson. The test statistic is

$$\mathrm{Sc}(\bar{y}, N, p_0) = \sqrt{\frac{N}{e^{\bar{y}} - 1 - \bar{y}}} \left(p_0 e^{\bar{y}} - 1\right), \tag{5.2.1}$$

where $p_0 = N_0/N$ and for $\mathrm{Sc}(\bar{y}, 0) > 1.645$, the Poisson hypothesis is rejected in favour of the ZIP model at 5% significance level. We note that this test is usually only considered for the purpose of cytogenetic-based dose estimation. More recently, Oliveira [76] developed a variant of (5.2.1) for use with the identity link, furthermore showing that the resulting test statistic was similar to that obtained under the log-link.

For the log-likelihood of a ZINB model without zero-inflation, $\ell(\mu_i, \alpha, 0)$, let us assume that the MLE of the population mean can be approximated by the mean $\bar{y}$ and the MLE of $\alpha$ is obtained maximising $\ell(\mu_i, \alpha, 0)$. Therefore the score evaluated under $H_0$, $(\bar{y}, \hat{\alpha}, 0)$, yields

$$\mathrm{Sc}(\bar{y}, \hat{\alpha}, 0) = \left(\frac{\partial \ell_0}{\partial \mu_i}, \frac{\partial \ell_0}{\partial \alpha}, 0\right)(\bar{y}, \hat{\alpha}, 0) = \left(0, 0, \frac{p_0}{\hat{\pi}_0(\bar{y}, \hat{\alpha})} - 1\right).$$

The corresponding Fisher information matrix is then calculated as the negative expectation of the second derivatives with respect to each parameter in $\theta$ of the log-likelihood of the sample. The second derivatives of $\ell_0$ with respect to $p$ are

$$\frac{\partial^2 \ell_0}{\partial p \partial \theta}(\mu_i, \alpha, p) = -N_0 \frac{p + (1-p)\pi_0(\mu_i, \alpha) + (1-p)(1 - \pi_0(\mu_i, \alpha))}{(p + (1-p)\pi_0(\mu_i, \alpha))^2} \frac{\partial \pi_0(\mu_i, \alpha)}{\partial \theta},$$

$$\frac{\partial^2 \ell_0}{\partial p^2}(\mu_i, \alpha, p) = -N_0 \frac{(1 - \pi_0(\mu_i, \alpha))^2}{(p + (1-p)\pi_0(\mu_i, \alpha))^2} - \frac{N - N_0}{(1-p)^2}.$$

The Fisher information under $H_0$ $(p = 0)$ is then

$$I(\theta_0) = I(\bar{y}, \hat{\alpha}, 0) = \begin{pmatrix} I(\bar{y}, \hat{\alpha}) & \frac{N}{\hat{\pi}_0(\bar{y}, \hat{\alpha})} \frac{\partial \hat{\pi}_0(\bar{y}, \hat{\alpha})}{\partial \mu_i} \\ & \frac{N}{\hat{\pi}_0(\bar{y}, \hat{\alpha})} \frac{\partial \hat{\pi}_0(\bar{y}, \hat{\alpha})}{\partial \alpha} \\ & \frac{N}{\hat{\pi}_0(\bar{y}, \hat{\alpha})} - N, \end{pmatrix}$$

where $I(,)$ is the information matrix of the model without zero-inflation. The score test for zero-inflation is given by

$$\begin{aligned} \mathrm{Sc}(\bar{y}, \hat{\alpha}, p_0) &= \mathrm{Var}(\hat{p}) \frac{\partial \ell_0}{\partial p}(\bar{y}, \hat{\alpha}, 0)^2 \\ &= \frac{\det(I(\bar{y}, \hat{\alpha}))}{\det(I_0(\bar{y}, \hat{\alpha}, 0))} \left( \frac{p_0}{\hat{\pi}_0(\bar{y}, \hat{\alpha})} - 1 \right)^2, \end{aligned} \tag{5.2.2}$$

where the asymptotic distribution under $H_0$ is a $\chi^2_1$. We are interested in a one-tailed test which can be done by using a signed version of (5.2.2) i.e.

$$\sqrt{\mathrm{Sc}(\bar{y}, \hat{\alpha}, p_0)} = \sqrt{\frac{\det(I(\bar{y}, \hat{\alpha}))}{\det(I_0(\bar{y}, \hat{\alpha}, 0))}} \left( \frac{p_0}{\hat{\pi}_0(\bar{y}, \hat{\alpha})} - 1 \right), \tag{5.2.3}$$

which is asymptotically distributed as a standard normal. A score test for a ZIP regression model against ZINB alternatives was introduced by [87]. The test is presented in a linear form and has a standard normal distribution if the model under the null hypothesis is true, however the computations are very complex and not compatible under the identity link.

## 5.3 Poisson vs NB

It is clear from the variance functions of both the NB1 and NB2 that they reduce to a Poisson when $\alpha = 0$. Therefore, testing the Poisson assumption against the NB alternative corresponds to testing:

$$H_0 : \alpha = 0 \qquad \text{against} \qquad H_1 : \alpha > 0$$

A likelihood ratio test or Wald test can be used, as conducted in Section 5.1, but the score test has the advantage that we only need to fit the Poisson model. The explicit forms for a score test were given by Dean [64]. For testing a Poisson regression against NB1 we have [77]:

$$\frac{1}{2N} \left( \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(y_{ij} - \hat{\mu}_i)^2 - y_{ij}}{\hat{\mu}_i} \right)^2$$

and for a Poisson against NB2:

$$\frac{(\sum_{i=1}^{k} \sum_{j=1}^{n_i} ((y_{ij} - \hat{\mu}_i)^2 - y_{ij}))^2}{2 \sum_{i=1}^{k} \hat{\mu}_i^{\,2}}$$

Under the hypothesis of the Poisson model, the limiting distribution of the score statistic is $\chi^2_1$. The test statistics are found to be 117.12 and 6446.02 for PHE-Foci1 and 12.18 and 7517.66 for PHE-Foci2, which means the null hypothesis is rejected

in favour of both NB models. Note that all three tests are asymptotically equivalent and all of the test statistics indicate evidence against the fit of the Poisson model to the data, yet we notice there is quite a difference in their realised values. A probable explanation is that the likelihood ratio and score tests are both strongly dependent on the sample size, more so for the latter, therefore using a relatively small sample may result in a larger test statistic.

## 5.4 Model selection

Often the presence or type of violation of the Poisson model will not always be obvious from the fitted calibration curve and if the response distribution is incorrectly specified, the uncertainty assessment will also be incorrect. In order to compare the performance of NB and zero-inflated models with the Poisson model, classical likelihood measures of goodness of fit can be used: the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The AIC [7] penalises a model with a larger number of parameters, and is defined as AIC$= -2\ell + 2w$. The BIC [98], defined as BIC $= -2\ell + w\log N$, works similarly to AIC but increases the penalty with increasing sample size. According to these criteria, models with smaller values of AIC and BIC are considered preferable.

| | | PHE-Foci1 | | | PHE-Foci2 | | |
|---|---|---|---|---|---|---|---|
| Model | $w$ | $-2\ell$ | AIC | BIC | $-2\ell$ | AIC | BIC |
| Poisson | 2 | 4175.61 | 4179.61 | 4189.79 | 14868.01 | 14872.01 | 14884.60 |
| NB1 | 3 | 4064.67 | 4070.67 | 4085.94 | 14858.72 | 14864.72 | 14883.60 |
| NB2 | 3 | 4126.10 | 4132.10 | 4147.37 | 14868.00 | 14874.00 | 14892.89 |
| ZIPa | 3 | 3968.85 | 3974.85 | 3990.12 | 14716.97 | 14722.97 | 14741.85 |
| ZIPb | 3 | 3837.17 | **3843.17** | **3858.44** | 14775.79 | 14781.79 | 14800.67 |
| ZIPc | 4 | **3836.49** | 3844.49 | 3864.85 | **14703.36** | **14711.36** | **14736.54** |
| ZINB1a | 4 | 3963.79 | 3971.79 | 3992.15 | 14716.88 | 14724.88 | 14750.06 |
| ZINB1b | 4 | 3837.23 | 3845.23 | 3865.60 | 14775.86 | 14783.86 | 14809.04 |
| ZINB1c | 5 | 3836.55 | 3846.55 | 3872.01 | 14703.41 | 14713.41 | 14744.88 |
| ZINB2a | 4 | 3968.85 | 3976.85 | 3997.21 | 14718.02 | 14726.02 | 14751.19 |
| ZINB2b | 4 | 3837.17 | 3845.17 | 3865.54 | 14825.89 | 14833.89 | 14859.06 |
| ZINB2c | 5 | 3837.79 | 3847.79 | 3873.24 | 14725.23 | 14735.23 | 14766.70 |

Table 5.1 Likelihoods and model criterion from fitting various models to 4h whole-body calibration data for datasets PHE-Foci1 and PHE-Foci2.

Table 5.1 show the values of the maximised log–likelihood as well as the information criteria, with the best model in each column provided in bold face. Firstly, we observe for the PHE-Foci1 dataset that the Poisson model provides the (by far) worst fit.

| Link | Test | PHE-Foci1 | | | PHE-Foci2 | | |
|---|---|---|---|---|---|---|---|
| | | $W$ | $R$ | Score | $W$ | $R$ | Score |
| Identity | P/ZIPa | 74.11 | 206.76 | 457.90 | 32.05 | 151.04 | 189.84 |
| | P/NB1 | 53.92 | 110.94 | 117.12 | 8.13 | 9.29 | 12.18 |
| | P/NB2 | 27.37 | 49.51 | 6446.02 | $< 10^{-3}$ | 0.01 | 7517.66 |
| | NB1/ZINB1a | 39.82 | 100.88 | 1527.24 | 31.28 | 141.84 | 2734.36 |
| | NB2/ZINB2a | 74.16 | 157.25 | 14406.41 | 30.79 | 149.98 | 15354.47 |
| | ZIPa/ZINB1a | 0.70 | 5.06 | 19.72 | $< 10^{-3}$ | 0.09 | 4476.84 |

Table 5.2 Results from the Wald, likelihood ratio and score tests. For testing ZIPa vs ZINB1a the score test is calculated under the log-link.

According to the log-likelihood, the ZIPc is considered the preferred model, however the ZIPb is selected by both the AIC and BIC. This is not concerning since the ZIPc is the second best of each criteria. By contrast, it appears that the NB2 and ZIPc (closely followed by the ZINB1c in terms of AIC and ZIPa for BIC) is the least and most adequate model respectively for the PHE-Foci2 dataset. Although our results in Table 5.2 from the NB vs ZINB tests indicate that the overdispersion in both datasets can mostly be explained by zero-inflation, confirming the p-values found in Chapter 3, it is preferable to utilise a ZINB model to account for any excess heterogeneity (albeit some evidence from the Wald and likelihood ratio tests suggesting suitability of the ZIPa model).

# Chapter 6

# The effect of data aggregation on dispersion estimates in count data models

For the modelling of count data, aggregation of the raw data over certain subgroups or predictor configurations is common practice. This is, for instance, the case for count data biomarkers of radiation exposure. Under the Poisson law, count data can be aggregated without loss of information on the Poisson parameter, which remains true if the Poisson assumption is relaxed towards quasi-Poisson. However, in biodosimetry in particular, but also beyond, the question of how the dispersion estimates for quasi-Poisson models behave under data aggregation have received little attention. Indeed, for real data sets featuring unexplained heterogeneities, dispersion estimates can increase strongly after aggregation, an effect which we will demonstrate and quantify explicitly for some scenarios. The increase in dispersion estimates also implies an inflation of the parameter standard errors, which, however, by comparison with random effect models, can be shown to serve a corrective purpose. The phenomena are illustrated for the BfS-Foci data, based on a smaller dose range. This chapter corresponds to the contents of [32].

## 6.1 Motivation

We have observed that both overdispersion (with the exception of underdispersion in the PHE-Foci2 3Gy sample) and zero-inflation are present in H2AX-foci data so that, for instance, quasi-Poisson or negative binomial models appear adequate. However,

a simple practical question arising is whether the model fitting can be carried out without loss of information using only the aggregated data, as displayed in Fig. 1.7, or whether the raw data, as exemplified in Fig. 1.6, should be used. This question is of greater depth than one would expect: While we demonstrate in the next Section, that, in theory, one would anticipate the dispersion to be unaffected by the aggregation, for the BfS-Foci dataset the dispersion estimate resulting from a quasi-Poisson fit using the raw data is 1.223, while the one resulting from the fit to the aggregated data is 147.99! In a further twist, we will also see that the inflated dispersion of the aggregated model is not necessarily useless: it is a manifestation of a problem which lies elsewhere, namely dependency structures within the raw data, and eventually leads to the estimation of parameter standard errors which are more correct than those of the raw data model. Even though the connection of data aggregation to overdispersion is not an unknown phenomenon (in fact, in the ecological literature, the term 'aggregated' is often used synonymous to 'overdispersed' [41]), we believe that the implications of count data aggregation on dispersion estimates and ensuing inferential purposes, are, so far, poorly appreciated in the biodosimetric community, and also lack explicit study in the statistical literature.

## 6.2 Fitting aggregated data models

With reference to the raw linear model (2.2.3), the corresponding aggregated (linear) model becomes

$$E(s_i|x_i) = \beta_0 \times n_i + \beta_1 \times (n_i x_i). \tag{6.2.1}$$

Under the assumption of independent foci counts, we anticipate the dispersion under the raw data model should, in theory, remain systematically the same following aggregation. After fitting the aggregated models, we note that the coefficients of the quasi-Poisson models do not change between the two data types. Hence, the calibration curves of expected foci yield given dose, as displayed in Fig. 6.1, will remain exactly the same if estimated through raw or aggregated data models. However, a significant difference is observed in their dispersions, where for the PHE-Foci1 data we obtain estimates of $\hat{\phi} = 1.444$ and $\hat{\phi}_{agg} = 37.70$ using the formulations presented in (4.1.2) and (4.1.4) respectively. For the BfS-Foci data, we observe a smaller raw dispersion of $\hat{\phi} = 1.223$ but a much larger aggregated dispersion of $\hat{\phi}_{agg} = 147.99$. We note that all these dispersions would lead to a rejection of the Poisson hypothesis with a $\chi^2$ goodness of-fit-test.

A possible source of increased dispersion for the aggregated data model as compared to the raw data model could be an increased variance of the estimates under the former. Immediately from (4.1.6), it is clear under the aggregated model, where $\nu$ is much smaller, $\text{Var}(\hat{\phi})$ is larger. Firstly, substituting $\hat{\phi} = 1.223$ in the right hand side of

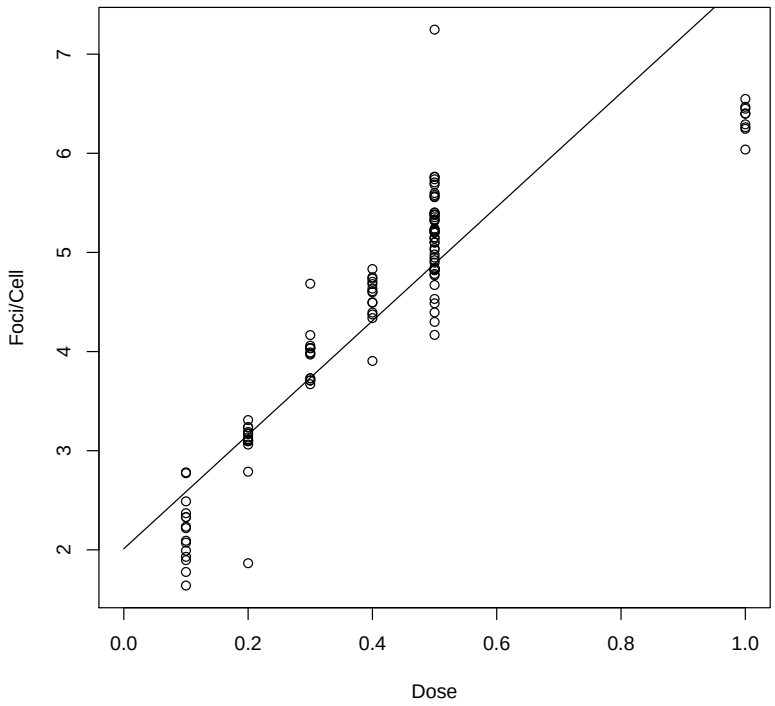



Fig. 6.1 Quasi-Poisson model estimates of the BfS-Foci linear calibration curve: $E(y_i) = 2.011 + 5.746x_i$.

(4.1.6), with degrees of freedom adjusted according to Table 6.1, leads to $\mathrm{SE}(\hat{\phi}) \approx 0.004$ for the raw model and $\mathrm{SE}(\hat{\phi}_{agg}) \approx 0.16$ for the aggregated model. However, this effect — to which we refer as *variance effect* henceforth — is certainly not sufficient to explain a value of, say, $\hat{\phi}_{agg} = 147.99$, for the dispersion of the linear fit to the aggregated data. We notice the standard error associated with the PHE-Foci1 aggregated dispersion is just the reported $\hat{\phi}$, which is simply a consequence of the aggregated data consisting of only $k = 4$ slides i.e. a slide per dose (in contrast to the BfS-Foci dataset which has $k = 116$ slides with multiple slides per dose). In such circumstances, it is difficult to infer the variance as the pinpoint cause for the increased dispersion without further data. Indeed, this highlights the importance of raw data being made available. For these reasons, we dedicate further attention to the BfS-Foci data in the sections which follow.

### 6.2.1 Random effect models

Since the data $y_{ij}$ do possess a two-level structure, with the slides $i$ corresponding to the upper level, and the foci frequencies within slides corresponding to the lower level, it appears adequate to contrast the previous results with an alternative modelling strategy where within-slide correlation is explicitly accounted for by an additive random effect,

| | Raw | | Aggregated | |
|---|---|---|---|---|
| | PHE-Foci1 | BfS-Foci | PHE-Foci1 | BfS-Foci |
| $(\hat{\beta}_0, \hat{\beta}_1)$ | (0.766, 1.700) | (2.011, 5.746) | (0.766, 1.700) | (2.011, 5.746) |
| $(\text{SE}(\hat{\beta}_0), \text{SE}(\hat{\beta}_1))$ | (0.042, 0.058) | (0.009, 0.023) | (0.213, 0.298) | (0.102, 0.248) |
| $\hat{\phi}$ | 1.444 | 1.223 | 37.70 | 147.99 |
| $\text{SE}[\hat{\phi}]$ | 0.049 | 0.004 | 37.70 | 0.16 |
| $\nu$ | 1198 | 233218 | 2 | 114 |
| $\chi^2_{\nu,0.95}/\nu$ | 1.068 | 1.005 | 2.996 | 1.227 |

Table 6.1 Parameter estimates along with their associated standard errors and dispersion estimates obtained from each model. The last row gives the critical value that $\hat{\phi}$ would be compared with in a Poisson goodness-of-fit test at the 5% level of significance.

also called random intercept, operating on the upper level. Hence, we consider a mean function of type $\tilde{\mu}_i = \mu_i + u_i$, where $\mu_i$ is as in (2.2.2), and $u_i \sim N(0, \sigma_r^2)$ is a Gaussian random effect. For the response distribution, we consider two scenarios, namely a Poisson mixed model $y_{ij} \sim \text{Pois}(\tilde{\mu}_i)$, and a NB1 regression model, $y_{ij} \sim \text{NB1}(\tilde{\mu}_i, \alpha)$ where $\phi = 1 + \alpha$. That is, the NB1 model allows the parameter $\phi$ to capture any dispersion not accounted for by the slide-wise random effect.

The models are fitted with R function `glmmTMB` [18], and results are provided in Table 6.2. We firstly observe that both models behave similarly, and that their standard errors lend, on comparison with Table 6.1, interestingly, support to the aggregated data model. This can be interpreted as that the dispersion estimate of the aggregated model has successfully captured the between-slide heterogeneity described by the random effect model. Informally, the presence of this heterogeneity is visible from the small but non-zero intra-class correlations (ICC). More formally, one can carry out statistical tests for the significance of the random effect term, with $H_0 : \sigma_r^2 = 0$. For the Poisson model, the likelihood ratio statistic of models with and without the random effect term is $2(513385.3 - 505138.5) = 16493.6$, clearly indicating rejection of $H_0$ when contrasting with a $0.5(\chi_0^2 + \chi_1^2)$ distribution. For the NB1 model, the conclusion is identical with $LR = 2(511119.5 - 503979.3) = 14280.4$. One can test for the significance of overdispersion ($H_0 : \phi = 1$) by comparing $\hat{\phi} = 1.141$ with $\chi^2_{0.95,233217}/233217 = 1.005$, also yielding significance. So, albeit just above 1, the value of 1.141 represents genuine overdispersion (over and above the one explained by the random effect model). In summary, this provides evidence of heterogeneities existing both between and within slides. It is furthermore noted that the coefficient estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ for the random effect model differ by about three standard errors from the raw and aggregated data models.

| | Mixed Poisson | Mixed NB1 |
|---|---|---|
| $(\hat{\beta}_0, \hat{\beta}_1)$ | (2.331, 4.974) | (2.327, 4.983) |
| $(\text{SE}(\hat{\beta}_0), \text{SE}(\hat{\beta}_1))$ | (0.114, 0.242) | (0.114, 0.243) |
| $\hat{\phi} = 1 + \hat{\alpha}$ | | 1.141 |
| $\hat{\sigma}_r^2$ | 0.334 | 0.337 |
| $\hat{\sigma}_\epsilon^2$ | 4.998 | 4.998 |
| ICC | 0.063 | 0.063 |

Table 6.2 Parameter estimates of the fitted random effect models. Results above the dashed line are extracted directly from the output of function `glmmTMB`. The values below the dashed line give the estimated residual variance, $\hat{\sigma}_\epsilon^2$, and the resulting ICC values.

### 6.2.2 Non-parametric bootstrap

Having seen the evidence for heterogeneities in the data, the models fitted in Section 6.2 can be considered misspecified. In order to understand better the impact of this misspecification on the fitted raw and aggregated data models, we carry out a bootstrap simulation, with the mixed NB1 model fitted in Section 6.2.1 as base model, and examine the dispersion estimates, and resulting standard errors, of all models.

The sampling process of this bootstrap is built in two stages (with all estimates taken from Table 6.2):

1. Generate slide-wise random errors $u_i^*$ by sampling from $N(0, \hat{\sigma}_r^2)$;
2. Simulate bootstrap data $y_{ij}^* \sim \text{NB1}(\hat{\beta}_0 + \hat{\beta}_1 x_i + u_i^*, \hat{\alpha})$.

Repeat 1. and 2. $B$ times to obtain $B$ bootstrap samples. Then, for each of the $B$ iterations, we fit three models:

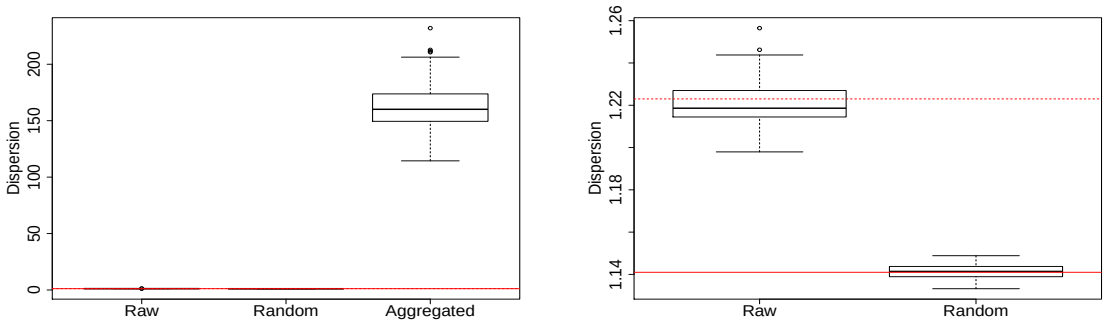
(i) A quasi-Poisson regression model with identity link, applied on the bootstrapped raw data $y_{ij}^*$, i.e. model (2.2.3).

(ii) A quasi-Poisson regression model with identity link, applied on the bootstrapped aggregated data $s_i^* = \sum_j y_{ij}^*$ i.e. model (6.2.1).

(iii) A NB1 regression model with identity link, applied on the bootstrapped raw data $y_{ij}^*$; with an additive random effect representing slides.

For each fitted model and bootstrap iteration, dispersion estimates for models (i) and (ii) are computed according to (4.1.2) and (4.1.4), respectively, with standard errors arising as explained in Section 4.1. For model (iii), this dispersion estimate is obtained by adding 1 to the 'overdispersion' parameter, $\hat{\alpha}$, reported in the summary

Fig. 6.2 Dispersion estimates based on the bootstrap simulation. The solid red line represents the random-effect model dispersion $\hat{\phi} = 1.141$ and the dashed line indicates the quasi-Poisson dispersion $\hat{\phi} = 1.223$ for the original data as reported in Tables 6.1 and 6.2.




output of R function `glmmTMB` [18]. Standard errors are extracted directly from this output.

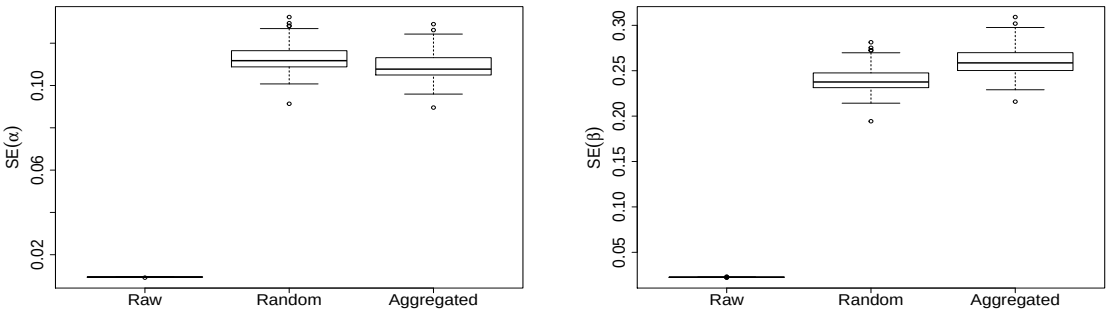
Boxplots of the dispersion estimates for the bootstrap simulation are displayed in Fig. 6.2. The left hand panel in this figure gives a comparison of the dispersion estimates for the raw, random, and aggregated models, whereas the right panel gives a zoomed comparison of the raw and random effect models. We see from this that the dispersion estimates for the raw data model are positioned close to the correct mean value at 1.223. However, the boxplot for the dispersion estimates from the aggregated model now sits at about 160, which is of similar magnitude as in our initial analysis in Table 6.1. While the variability of these estimates is also larger than for the raw data model, it is clear that something much more drastic (than just inflation of variance) has occurred here, shifting the bulk of the dispersion estimates from the magnitude 1-2 to much larger values. The dispersion estimates from the random effect model are slightly smaller than for the raw data model, centering correctly at the value 1.141 from which the data were generated, as visible from the right panel. The slight difference between these two models is plausible, as some of the original overdispersion has been captured by the random effect.

| | Raw | Random | Aggregated |
|---|---|---|---|
| $\text{SD}(\hat{\beta}_0)$ | 0.109 | 0.103 | 0.109 |
| $\text{SD}(\hat{\beta}_1)$ | 0.226 | 0.211 | 0.226 |

Table 6.3 Mean parameter standard deviations based on 100 simulation runs.

We investigate now the consequences of this inflated dispersion. Therefore, let us firstly consider the boxplots in Fig 6.3. It is clear from this that, for the raw data model, the reported standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are very small. However, we deduced from

Fig. 6.3 Parameter standard errors for the bootstrap simulation (left: intercept; right: slope).




Tables 6.1 and 6.2 that either of aggregation, or the use of a random effect, transports the standard errors to a much higher level. The values reported in Table 6.3 reveal that, over all estimation methods, the actual *standard deviation* of the bootstrapped estimates of regression coefficients is very similar, and is *for all three models, including the raw data model*, of the (high) magnitude reported by the aggregated and random effect models. This, in turn, implies that the standard errors of regression parameters for the raw data model, as reported in Table 6.1 and Fig. 6.3, are wrong. We arrive, hence, at the intriguing conclusion that the large dispersion produced by the aggregated data model serves eventually a good purpose — namely to adjust the standard errors of the parameter estimates so that these match the magnitude of those from the random effect model. For later reference, we will refer to this effect, i.e. the tendency of aggregated data models to inflate dispersion estimates in order to account for violations of the independence assumption in the raw data, as a *dependency effect*.

## 6.3 Special case: mixture-induced heterogeneity

In this section we will make the "dependency effect" more explicit by mathematically deriving the inflation factors for an important special case: The case of a mixture model without covariates.

### 6.3.1 A two-component model inducing heterogeneity

Consider a scenario in which we generate $k$ rows (slides), each consisting of $n_i \equiv n$ Poisson foci counts (cells), but for fixed covariate dose (in other words, in the absence of covariates). However, we assume that there exists heterogeneity, that is some counts are from a Pois($\lambda_1$) distribution with probability $q$ (the Bernoulli parameter which selects the Poisson mean) and others from a Pois($\lambda_2$) with probability $1 - q$. In general

terms, the Poisson means come from a two-point mixture; i.e. each raw count $y_{ij}$ is generated as

$$y_{ij} \sim Z_{ij}\text{Pois}(\lambda_1) + (1 - Z_{ij})\text{Pois}(\lambda_2) \tag{6.3.1}$$

where $Z_{ij} \sim B(1, q)$. The resulting heterogeneity creates overdispersion which, under model (6.3.1), can be exactly quantified as

$$\phi = \frac{\text{Var}(y_{ij})}{E(y_{ij})} = 1 + \frac{q(1-q)(\lambda_1 - \lambda_2)^2}{q\lambda_1 + (1-q)\lambda_2}. \tag{6.3.2}$$

See Appendix A.7.2 for proof of this statement and Fig. 6.4 (top) for a visual representation of $\phi$ as a function of $q$; note also that the dependence on $x_i$ is now suppressed as there are no covariates. Expression (6.3.2) holds true even if there are correlation structures within the $Z_{ij}$. However, we will see that, for the dispersion of the aggregated data, it makes a crucial difference whether the heterogeneity is entirely random (i.e. the indicators $Z_{ij}$ are independently generated for all $i$ and $j$), or whether there is some correlation structure.

Consider, for instance, a scenario in which

$$Z_{ij} \equiv Z_i \qquad \text{for all} \quad j = 1, \ldots, n, \tag{6.3.3}$$

that is all counts within each slide are generated from a Poisson distribution with the same mean, but there is 2-component heterogeneity between slides. Then, one finds for $j \neq l$ by the law of total covariance,

$$\begin{aligned} \text{Cov}(y_{ij}, y_{il}) &= E(\text{Cov}(y_{ij}, y_{il})|Z_i) + \text{Cov}(E(y_{ij}|Z_i), E(y_{il}|Z_i)) \\ &= \lambda_1^2\text{Var}(Z_i) + \lambda_2^2\text{Var}(1 - Z_i) + 2\lambda_1\lambda_2\text{Cov}(Z_i, 1 - Z_i) \\ &= q(1-q)(\lambda_1 - \lambda_2)^2 \end{aligned} \tag{6.3.4}$$

so for $\lambda_1 \neq \lambda_2$ the independence assumption in (*) in (4.1.3), Section 4.1 is clearly violated. Depending on the mechanism generating the $Z_{ij}$, this expression will look different, but the point is that any dependency structures within the $Z_{ij}$ will render these covariances non-zero.

### 6.3.2 Theoretical dispersion of aggregated data

Aggregated data are obtained as before as $s_i = \sum_{j=1}^{n} y_{ij}$. The object of interest in this subsection is $\phi_{agg} = \text{Var}(s_i)/E(s_i)$, where we have now made notationally explicit that it may be different from $\phi$. Through the law of total expectation and variance one can

show that (see Appendix A.7.2), under model (6.3.1)

$$E(s_i) = n(q\lambda_1 + (1-q)\lambda_2); \tag{6.3.5}$$

$$\mathrm{Var}(s_i) = n(q\lambda_1 + (1-q)\lambda_2) + nq(1-q)(\lambda_1 - \lambda_2)^2 + \sum_{j\neq l}^{n} \mathrm{Cov}(y_{ij}, y_{il}). \tag{6.3.6}$$

This gives a general expression for the aggregated dispersion,

$$\begin{aligned} \phi_{agg} &\equiv \frac{\mathrm{Var}(s_i)}{E(s_i)} \\ &= 1 + \frac{q(1-q)(\lambda_1-\lambda_2)^2}{q\lambda_1 + (1-q)\lambda_2} + \frac{\sum_{j\neq l}^{n} \mathrm{Cov}(y_{ij}, y_{il})}{n(q\lambda_1 + (1-q)\lambda_2)}. \end{aligned} \tag{6.3.7}$$

In the simplest case that all covariances are identical to 0, the third term disappears and one sees immediately that $\phi_{agg}$ corresponds to the expression for $\phi$ given in (6.3.2). In the previously discussed case of slide-wise dependencies (6.3.3), one finds by using expression (6.3.4) and then referring to (6.3.2) that

$$\phi_{agg} = 1 + \frac{nq(1-q)(\lambda_1-\lambda_2)^2}{q\lambda_1 + (1-q)\lambda_2} = 1 + n\,(\phi - 1)\,. \tag{6.3.8}$$

We discuss a third scenario which we consider of practical relevance. Assume there are correlated strings of length $\tau < n$, each sharing the same Poisson mean. One can consider this as a special case of model (6.3.1) where the indicators $Z_{ij}$ share the same value for blocks of length $\tau < n$, in terms of the index $j$. Then one can show (Appendix A.5.2) that

$$\phi_{agg} = 1 + \tau\,(\phi - 1)\,, \tag{6.3.9}$$

neatly extending (6.3.8). Note that both $\phi$ and $\phi_{agg}$ can be considered as functions of the mixing proportion, $q$. This is visualised in Fig. 6.4 (bottom). We take note of the non-symmetry in terms of the mixing parameter, with a maximum at $q = 2 - \sqrt{2}$. Furthermore, we observe that for $q = 0$ or $q = 1$ there is no overdispersion since there is no heterogeneity.

Equations (6.3.8) and (6.3.9) provide some insight into how the presence of different types of heterogeneity, for example through correlation within rows or strings within rows, affect the dispersion of the aggregated data. From direct inspection of both (6.3.8) and (6.3.9), we deduce that if one increases either the row length or the string size then the dispersion of the aggregated data continues to grow larger. We also notice that if there is no overdispersion of the raw counts, i.e. $\phi = 1$, then we have equidispersion for the aggregated data as expected. If one has only clusters of size 1 ($\tau = 1$ or $n = 1$; that

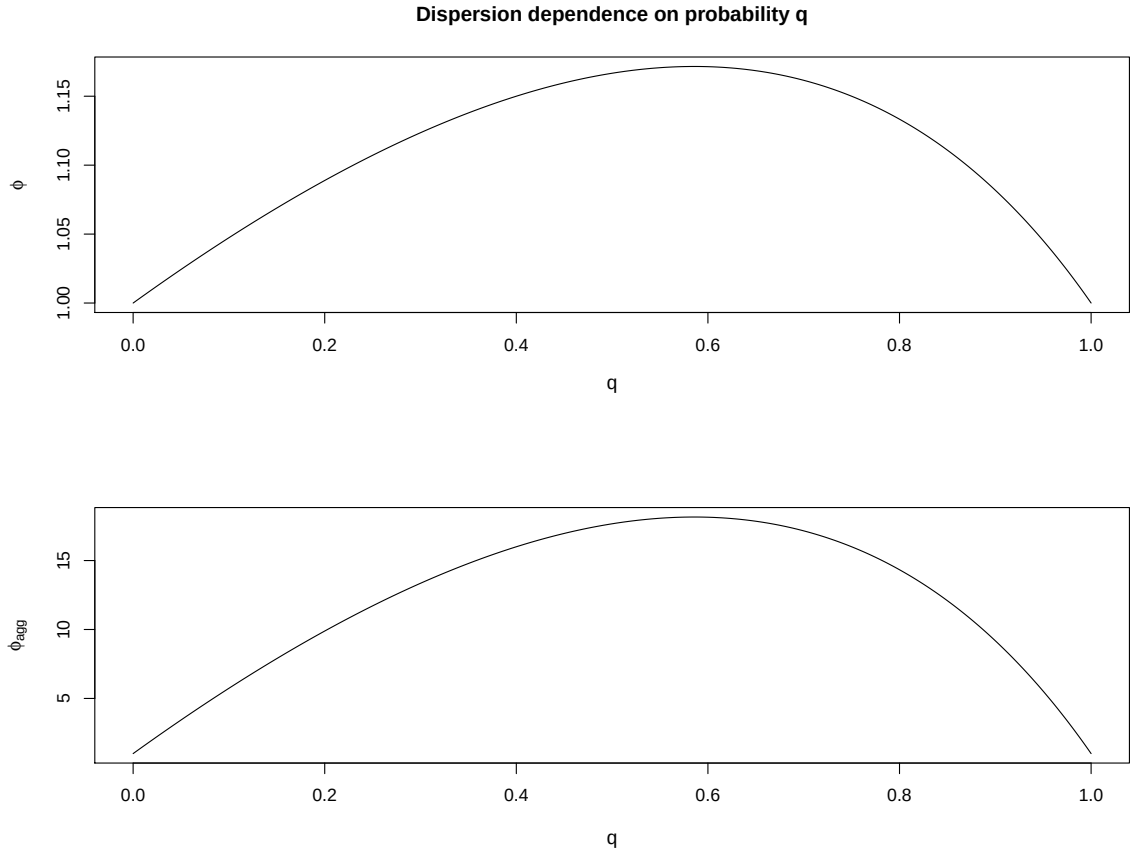



Fig. 6.4 For fixed $\lambda_1 = 1, \lambda_2 = 2$, we plot the non-linear functions (6.3.2) and (6.3.9), using a string size of $\tau = 100$. Note the substantially different scales in the vertical axes of the two plots.

is, the heterogeneity is entirely random) then $\phi_{agg} = \phi$, so in this case the aggregated data dispersion does not inflate.

### 6.3.3 Experiment

We carry out a simulation experiment as described in Section 6.3.1 using $\lambda_1 = 1$, $\lambda_2 = 2$ and $q = 0.5$. The mechanisms presented in Section 6.3.1 and the theoretical derivations in Section 6.3.2 mean that the heterogeneity resulting from the mixture will trigger overdispersion, but that the overdispersion for the aggregated data will depend on the correlation structure of the heterogeneity-inducing mechanism. This leads us to distinguish the following three cases:

(A) Random heterogeneity: For each slide and cell, the $Z_{ij}$ in (6.3.1) are generated independently;

(B) Slide-wise heterogeneity: The $Z_{ij}$ are generated once for each slide and kept constant for all cells in that slide, i.e. $Z_{ij} = Z_i$ as in (6.3.3);

(C) String-wise heterogeneity: The $Z_{ij}$ share the same value for blocks of size $\tau = 100$ within each slide, but different blocks are generated independently.

For each of (A), (B) and (C), $k = 1000$ slides of length $n = 1000$ are generated. Since no covariates are involved in this study, we do not need to fit any models to estimate dispersion. For the raw data, the dispersion is estimated by the overall dispersion index (3.1.1). For the aggregated data, this would be replaced by $\sum_{i=1}^{k}(s_i - \bar{s})^2/[(k-1)\bar{s}]$. The resulting dispersion values are reported in Table 6.4, with corresponding R code detailed in Appendix A.7.1.

| | (A) | (B) | (C) |
|---|---|---|---|
| Raw data | 1.167 | 1.168 | 1.166 |
| Aggregated data | 1.070 | 168.97 | 16.85 |

Table 6.4 Dispersion indexes from simulated data under scenarios (A), (B) and (C).

We can see that for case (A) the dispersion of the aggregated data does not increase at all, while in (B) we observe the strongest inflation. To reiterate, “aggregated data" signifies here row-wise (slide-wise) sums. Our BfS-Foci dataset best corresponds to (C) rather than (B), although the basis of the effect is the same.

Verifying these results through our theoretical derivations from Section 6.3.2, one obtains for case (B) via (6.3.8) that

$$\phi_{agg} = 1 + 1000\,(1.168 - 1) = 169,$$

which agrees closely with the simulated value of 168.97. Under scenario (C), where slides are split into 10 clusters each containing 100 cells ($\tau = 100$), one gets from (6.3.9)

$$\phi_{agg} = 1 + 100\,(1.166 - 1) = 17.6,$$

again in reasonable agreement with our simulation result of 16.85.

### 6.3.4 3-component Poisson mixture

Let us assume instead that some observations are from a Pois($\lambda_1$) distribution with probability $q_1$, some from a Pois($\lambda_2$) with probability $q_2$ while others are from a Pois($\lambda_3$) with probability $q_3 = 1 - q_1 - q_2$. Each raw count $y_{ij}$ is generated as

$$y_{ij} \sim \sum_{m=1}^{3} Z_{ijm}\text{Pois}(\lambda_m),$$

where $Z_{ijm} \sim B(1, q_m)$. The over-dispersion in this case is given by:

$$\begin{aligned}\phi &= \frac{\mathrm{Var}(y_{ij})}{E(y_{ij})} \\ &= 1 + \frac{\sum_{m=1}^{3} q_m(1-q_m)\lambda_m - 2q_1q_2\lambda_1\lambda_2 - 2q_1q_3\lambda_1\lambda_3 - 2q_2q_3\lambda_2\lambda_3}{\sum_{m=1}^{3} q_m\lambda_m}.\end{aligned}$$

For the aggregated data, defined before as $s_i = \sum_{j=1}^{n} y_{ij}$, one arrives at the same expressions as in (6.3.8) and (6.3.9). The resulting dispersion values corresponding to the three heterogeneity scenarios with $\lambda_1 = 1$, $\lambda_2 = 2$ and $\lambda_3 = 3$ and equal probabilities i.e. $q_1 = q_2 = q_3 = 1/3$ are reported below in Table 6.5. For comparison, under case (B) with $n = 1000$ one obtains from (6.3.8) that

$$\phi_{agg} = 1 + 1000\,(1.328 - 1) = 328$$

and for scenario (C) with $\tau = 100$ in (6.3.9)

$$\phi_{agg} = 1 + 100\,(1.335 - 1) = 34.5$$

therefore in fairly good agreement.

| | (A) | (B) | (C) |
|---|---|---|---|
| Raw data | 1.334 | 1.328 | 1.335 |
| Aggregated data | 1.326 | 332.38 | 33.04 |

Table 6.5 Dispersion indexes from simulated data under simulation scenarios as described in Section 6.3.3.

### 6.3.5 Generalisation of the model

We have provided this analysis for a 2-component mixture. Even if this constitutes a gross simplification of reality, we believe that this scenario represents the character of the phenomenon accurately. To underline this point, we included in the previous section a corresponding analysis for a 3-component mixture. In practice, and especially for our data, counts are likely to originate from more than two or three Poissons, however we do not expect the results to change in substance under a mixture of $M \geq 3$ Poisson random variables.

As the Multinomial model is often not suitable when there is observed overdispersion, the Dirichlet multinomial distribution model can be used as an alternative [54]. For further consideration, one may consider a model of type $y_{ij} = \sum_m Z_{ijm}\mathrm{Pois}(\lambda_m)$ where

variations among the component probabilities $q_m = P(Z_{ijm} = 1)$ follow a Dirichlet distribution, i.e. $q_m \sim \text{Dir}(\alpha)$, $m = 1, \ldots, M$, indicating that $y_{ij}$ belongs to component $m$ with probability $q_m$.

## 6.4 Summary for the practitioner

In many applied sciences, the use of aggregated count data is common, since they contain all relevant information to estimate Poisson models. Aggregated data are also usually less expensive to store and analyse than individual data. Another reason for the use of aggregated data is just convenience: While for biomarkers based on chromosomal aberrations, such as the dicentric assay, where counts larger than 7 or 8 are rarely observed, the full count distributions can still be conveniently displayed [76], this is is not necessarily the case for H2AX foci data where this count may be much higher. The data analyst may never get to see the raw data, and then has to work with the aggregated data simply as this is all that is available to them [28].

Under the presence of overdispersion, a conditional independence assumption of the responses given covariates guarantees, in theory, equality of the raw and aggregated data dispersion. However, we have seen that dispersion estimates for raw and aggregated data can differ dramatically for practical data sets. We distinguished that there are two effects which jointly result in an increased dispersion for the aggregated data model; a (relatively minor, but still significant) variance effect and a (potentially huge) dependency effect. We have demonstrated the latter phenomenon via example, simulation, and theory, uncovering in this process that the causes for the dependency effect reside in correlations between or within the slides being aggregated over. Another way of putting these findings is: The presence of unobserved heterogeneity will cause overdispersion in the raw data. If this heterogeneity follows dependency patterns (within or between slides), then this will lead to *inflated* overdispersion for the aggregated data. While the theoretical derivations, in Section 6.3, only cover the covariate-free case, they still give useful insights into the relationship of raw and aggregated data dispersion; specifically the aggregated dispersion increases linearly with the length of correlated strings within the data set, attaining a maximum if the string size corresponds to the full slides.

The relevant question is then whether the raw data should have been used if they were available, and if so, using which model. Under the presence of, say, slide-wise correlations in the raw data, the statistically sound model would be the use of a mixed model for the raw data which features a random intercept for each slide. It appears that such a model produces roughly similar parameter standard errors than the aggregated data model, whereas the raw data model produces much smaller standard errors. This appears to indicate that the high dispersion produced by the aggregated model is an

attempt by the model to solve a problem which resides somewhere else (namely in the between-slide-correlations), which the raw data model is not able to address (without the inclusion of random effects). Putting it into other words, the aggregated model finds a way to produce roughly *correct* uncertainty quantification by using *incorrect* dispersion estimates.

Random effects, however, have some practical limitations. For H2AX data, the main drawback of utilising a model with slide-specific random effect is that this random effect would be unknown for a newly exposed individual, which constitutes a major limitation as far as dosimetry is concerned. Furthermore, they can only account for between-slide correlations, but not within-slide correlations.

A practical advice to laboratories is to reduce heterogeneities to an absolute minimum, as they inflate dispersions and standard errors, and may also shift the actual calibration curve parameters. We do advise against using raw data models without adjustment by a random effect, but we do not advise against using the aggregated data models. Aggregation on the slide level does account for the correlations just as the random effect model would do, albeit using a much simpler model. On the contrary, our findings appear to justify their use in dosimetry to some extent, due to the implicit correction of standard errors. However the data analyst should be aware that the resulting dispersion estimates may be far from the underlying true dispersion of the raw data. This is of particular importance with view to the detection of partial body exposures through dispersion estimates, as is a common approach for dicentric chromosomes [47]. Inflated dispersions of the magnitude as observed would certainly render any attempt at identifying partial body exposure ineffective, unless one finds a way of working backwards to recover the raw data dispersion, for instance using equations such as derived in Section 6.3.

For known dispersions $\hat{\phi} > 1$ and $\hat{\phi}_{agg}$, one can estimate the size of the correlated sub-strings within each slide by rearranging (6.3.9) such that

$$\tau \approx \frac{\hat{\phi}_{agg} - 1}{\hat{\phi} - 1}. \tag{6.4.1}$$

For the BfS dispersions of $\hat{\phi} \approx 1.2$ and $\hat{\phi}_{agg} \approx 150$, we have correlated sub-strings of size $\tau = 745$ within each slide of $n \approx 2000$. For $\hat{\phi} < 2$ (as is commonly the case for H2AX), it is clear from (6.4.1) that $\tau > \hat{\phi}_{agg} - 1$. To the best of our knowledge, we are concerned only with values of $\tau$ based on $1 < \hat{\phi} \leq 1.5$ (i.e. small overdispersion). For multilevel H2AX data, the ICC (as calculated in Table 6.2) corresponds exactly to the correlation of two cells drawn randomly from the same slide. For randomly drawn cells from different slides, this correlation would be zero. Further work would be required to seek a possible relation between $\tau$ and the correlation coefficient (or ICC).

# Chapter 7

# The contaminated negative binomial method for estimation of radiation dose and exposure fraction

Many situations can be envisaged for which radiation casualties would only result in partial exposures to radiation. In this type of situation, even if personal dose meters are available or physical dosimetry methods [3], the intrinsic localisation of these dosimeters means that total body exposures may be vastly under- or over-estimated. Methods have been developed to adapt the dicentric assay for detection of partial body exposures. The IAEA manual recommends either the contaminated Poisson method or the QDR method. For estimating the irradiated fraction size, it has been reported that the QDR approach was generally found to be less accurate than the contaminated Poisson algorithm [97]. While an attempt to apply the contaminated Poisson to H2AX data has been made, it still remains that there are no methods to potentially allow for the detection of partial body exposures for biomarkers which *per se* produce overdispersed count distributions. In any partial exposure scenario, it is important to correctly quantify the fraction of exposure as otherwise the resulting dose estimates will also be incorrect thereby leading to potentially severe consequences for the exposed individual.

## 7.1 Proposed methodology

In Chapter 4 we considered the analysis of calibration data, in particular the purpose of whole-body calibration data for estimation of dose-response curves. However, in practice, following a potential radiation incident, a clinician or practitioner will usually

only be provided with a single exposed patient's blood sample for examination, in which a reference laboratory whole-body generated curve would then be used (in conjunction with the contaminated Poisson) to determine the contracted radiation dose and fraction of exposure. Here, we will make use of our own estimated Poisson/quasi-Poisson curves. We note that a calibration curve is needed as a preliminary to the considerations which follow.

### 7.1.1 The contaminated Poisson

For dose estimation, if data are consistent with a Poisson distribution then the recommendation is to report an averaged whole-body estimate. For unpicking a part-body exposure in the case of an overdispersed distribution of aberration counts, the current procedure as outlined in the IAEA manual is to follow Dolphin's method also named the "Contaminated Poisson" (CP) method. This method, which can be derived from the score equations of a ZIP, considers that the observed overdispersed distribution of counts can be expressed as the sum of two components;

1. A Poisson distribution representing the irradiated part of the body and
2. the remaining unexposed, and hence undamaged, fraction.

The following expression can be used to estimate a value $\hat{\mu}$ representing the yield of the irradiated part which is then substituted into a calibration curve (2.2.3) to estimate the contracted dose [26, 47]:

$$\frac{\hat{\mu}}{1-e^{-\hat{\mu}}} = \frac{s^*}{n^* - n_0^*} \tag{7.1.1}$$

where $e^{-\hat{\mu}}$ indicates the expected number of undamaged cells in the irradiated fraction according to the Poisson model. The left hand side of (7.1.1) represents the expectation of the zero-truncated Poisson distribution. We note that (7.1.1) is simply a rearranged version of (4.3.7), here updated for the case of a single inhomogeneous sample using the notation described in Section 2.3. An estimate for the variance of $\hat{\mu}$ can be obtained using the following expression (see Appendix A.4 for the derivation)

$$\mathrm{Var}(\hat{\mu}) = \frac{n_0\hat{\mu}^2(1-e^{-\hat{\mu}}))}{n\bar{y}\left[(1-e^{-\hat{\mu}})(n_0 - n(\hat{\mu}-\bar{y})e^{-\hat{\mu}}) - n\hat{\mu}e^{-2\hat{\mu}}\right]},$$

where the 95% confidence limits for both $\hat{\mu}$ and $\hat{p}$ (modelled via (4.4.12)) can be found in the usual way,

$$\hat{\mu}_{U/L} = \hat{\mu} \pm 1.96\sqrt{\mathrm{Var}(\hat{\mu})},$$

$$\hat{p}_{U/L} = \hat{p} \pm 1.96\sqrt{\text{Var}(\hat{p})}.$$

In the partial exposure scenario in which 25% blood has been exposed, it is clear that the remaining 75% will contribute very little foci. There does exist a background prevalence, for instance caused by naturally occurring ionising radiation, but this rate is usually considered to be very low which contradicts our motivations in later sections and certainly the following chapter. Hence, one naturally would assume that 75% of the sample consists of 'structural' zeros. However, even after accounting for zero-inflation, the overdispersion resulting from experimental factors cannot always be removed. We therefore require an alternative procedure which can be used to cover overdispersion arising from excess zeros and other sources.

### 7.1.2 The contaminated negative Binomial

The deficiencies of the contaminated Poisson method may be overcome by assuming that the scored foci in the irradiated fraction follow a negative binomial distribution. Recalling from (4.3.11), the MLE of $\mu$ for the ZINB1 can be found solving numerically

$$\hat{\mu}(n^* - n_0^*) = s^* \left(1 - (1 + \hat{\alpha})^{\frac{-\hat{\mu}}{\hat{\alpha}}}\right), \tag{7.1.2}$$

which we refer to as the "Contaminated NB" or "CNB". Unlike for the CP method, there does not exist simplifed forms for $\text{Var}(\hat{\mu})$ and $\text{Var}(\hat{p})$ but these quantities can be estimated using the code provided in Appendix A.4. In Section 7.2 we will attempt to showcase through simulated Poisson and overdispersed H2AX data the differences in dose and fraction estimates between the CP and the proposed CNB method, before applying these techniques to our practical data.

### 7.1.3 Estimation step

Given a blood sample consisting of $n^*$ foci counts observed under a certain level of dose and known "whole-body" calibration curve as in (2.2.3), of which can be validated through [28], the task is to arrive at an estimate for the absorbed dose $D$ and exposed fraction $F$. For the CNB method motivated by (7.1.2), the steps of the procedure are as follows:

S1 Compute the dispersion parameter $\alpha$ by fitting a ZINB1 regression (with identity link) to the sample.

S2 Extract the zero-inflation parameter estimate, $\hat{p}_{ZINB1}$, from the model output.

S3 The corresponding fraction $F_{CNB}$ can be estimated using (4.3.15).

S4 Given $s^*$ and $n_0^*$ (and $\hat{\alpha}$), solve (7.1.2) to obtain an estimate for $\mu_{CNB}$.

S5 One can now proceed to find a dose estimate $\hat{D}_{CNB}$ by replacing the sample mean, $\lambda$, with $\hat{\mu}_{CNB}$ in the dose response-curve (2.2.3). The linear case of the calibration curve motivates the dose estimator

$$\hat{D} = \frac{\hat{\mu}_{CNB} - \hat{\beta}_0}{\hat{\beta}_1}. \tag{7.1.3}$$

In the event that $\hat{\beta}_0 > \hat{\mu}_{CNB}$ (i.e a negative dose is produced), then $\hat{D}_{CNB}$ is set to zero.

For the calculation of $\hat{F}_{CP}$ and $\hat{D}_{CP}$ in the CP method, we begin from S2 (ZIP model is replaced in S2 for $\hat{p}_{ZIP}$ and instead we solve (7.1.1) for $\hat{\mu}_{CP}$).

### 7.1.4 Method of moments

To complement the previous section, the moments-based estimation offers a model-free approach and makes complete use of the sample summary statistics by using the relation:

$$\hat{\mu}_M = \frac{y^*}{1 - \frac{n^*}{n_0^*}},$$

which leads to

$$\hat{D}_M = \frac{\hat{\mu}_M - \hat{\beta}_0}{\hat{\beta}_1} \qquad \text{and} \qquad \hat{F}_M = 1 - \frac{n^*}{n_0^*} = \frac{y^*}{\hat{\mu}_M}.$$

### 7.1.5 Uncertainty under overdispersion

We recall that the uncertainty methods outlined in Section 2.4 are for the purpose of whole-body exposures, where it is assumed that the sampling data are Poisson. We have gathered sufficient evidence to backup the claim that $\gamma$-H2AX data (and in some cases dicentric data) no longer conform to the Poisson in both partial and whole-body exposures, therefore continuing with but failing to update these standard methods will likely to lead to grossly underestimated uncertainties. Both the delta and MC methods are advantageous in that they consider the sampling distribution via the error associated with the yield. Normally one would use the Poisson sampling error (PSE) for $\mathrm{SE}(y^*)$ in (2.4.1), however it is reasonable to replace this with either the QPSE or $\sqrt{\mathrm{Var}(\hat{\mu})}$ as calculated separately for the CP and CNB.

To represent a more rigorous approach with Merkle's method (and hence the simplification method), it has been suggested to use a correction factor for confidence

intervals of overdispersed WBI data. Given a sample with mean $y^*$ and dispersion index $\hat{\delta}$ (note that this is equivalent to the QP dispersion from an intercept-only model), the limits in (2.4.2) should be adjusted according to [3]

$$\tilde{\hat{\mu}}_{U/L} = \hat{\mu}_{U/L} \left( \frac{\hat{\mu}_{U/L}}{y^*} \right)^{\sqrt{\hat{\delta}}}.$$

In the case of PBI there remains no adjustments, however we propose to use the expression in Section 7.1.1 to calculate the confidence limits $\hat{\mu}_{U/L}$ as opposed to (2.4.2) since it allows Merkle's and the simplification method to capture the yield uncertainty.

## 7.2 Simulation study

In the context of dicentrics, assumed to be Poisson, the contaminated Poisson can be used to infer the degree of partial body exposure from the overdispersion. However, experimental factors which contribute variability in the scoring process of cells are all absorbed by the dispersion value. Although we have seen our datasets are heavily zero-inflated, such dispersion-generating effects cannot simply be ignored. In this section we aim to showcase the effect of disentangling dispersion from zero-inflation on both dose and fraction estimates.

In order to simulate H2AX-type foci count samples, we make use of an existing whole-body calibration curve reported in the literature [94]:

$$\lambda = 0.35(\pm 0.26) + 1.48(\pm 0.26)D. \tag{7.2.1}$$

Assuming a fixed dose $D$ for this simulation, $n^* = 1000$ observations were taken separately from two scenarios:

**A.** Poi($\lambda$)

**B.** NB1($\lambda$; $\alpha = 1$)

with 'base' dispersion $\phi = 1 + \alpha$, providing an equidispersed (A) and an overdispersed (B) whole-body sample (as exemplified in Figure 7.1). By "base dispersion" we refer to dispersion arising from possible bias in the sampling (or scoring) procedure and not from structural zeros due to some underlying physical reason. Additionally, we would like to showcase specifically how the CNB method can be used to disentangle dispersion and zero-inflation which will arise as a result of partial-body exposures. To mimic the 75% and 50% partial exposure scenarios, a proportion (corresponding to the fraction of exposure) of observations were randomly removed and then zeros were added to these samples. Hereafter, we make the assumption that information regarding

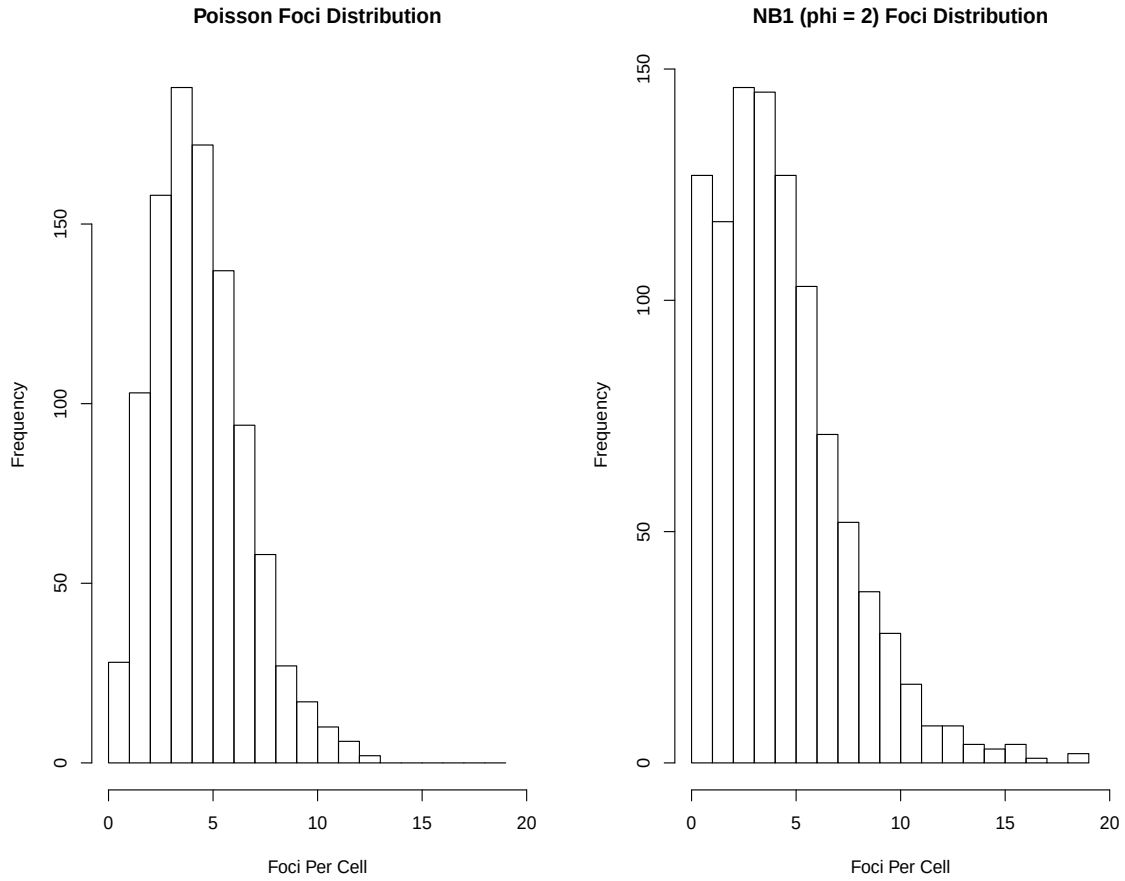



Fig. 7.1 A comparison of the individual number of foci per cell produced in equidispersed (left) and overdispersed (right) whole-body samples of equal mean.

| $F$ (%) | Dose $D$ (Gy) | $\hat{\mu}_{CP}$ | $\hat{\mu}_{CNB}$ | SD($\hat{\mu}_{CP}$) | SD($\hat{\mu}_{CP}$) | $\hat{\alpha}$ | $\hat{D}_{CP} \pm$ SD | $\hat{D}_{CNB} \pm$ SD |
|---|---|---|---|---|---|---|---|---|
| 75 | 1 | 1.817 | 1.801 | 0.059 | 0.064 | 0.032 | $0.991 \pm 0.040$ | $0.980 \pm 0.043$ |
| | | ($\pm 0.037, \pm 0.045, \pm 0.061$) | $\pm 0.066$ | | | | | |
| | 3 | 4.806 | 4.805 | 0.087 | 0.088 | 0.015 | $3.011 \pm 0.059$ | $3.001 \pm 0.059$ |
| | | ($\pm 0.060, \pm 0.089, \pm 0.082$) | $\pm 0.088$ | | | | | |
| | 5 | 7.752 | 7.752 | 0.095 | 0.095 | 0.020 | $5.001 \pm 0.064$ | $5.001 \pm 0.064$ |
| | | ($\pm 0.076, \pm 0.131, \pm 0.102$) | $\pm 0.099$ | | | | | |
| 50 | 1 | 1.822 | 1.802 | 0.076 | 0.077 | 0.040 | $0.994 \pm 0.051$ | $0.981 \pm 0.052$ |
| | | ($\pm 0.030, \pm 0.042, \pm 0.075$) | $\pm 0.044$ | | | | | |
| | 3 | 4.793 | 4.791 | 0.109 | 0.110 | 0.022 | $3.002 \pm 0.074$ | $3.001 \pm 0.074$ |
| | | ($\pm 0.049, \pm 0.090, \pm 0.100$) | $\pm 0.091$ | | | | | |
| | 5 | 7.756 | 7.756 | 0.108 | 0.108 | 0.029 | $5.004 \pm 0.073$ | $5.004 \pm 0.073$ |
| | | ($\pm 0.062, \pm 0.138, \pm 0.125$) | $\pm 0.124$ | | | | | |
| 75 | 1 | 2.282 | 1.832 | 0.092 | 0.131 | 0.982 | $1.306 \pm 0.062$ | $1.001 \pm 0.088$ |
| | | ($\pm 0.037, \pm 0.058, \pm 0.072$) | $\pm 0.063$ | | | | | |
| | 3 | 4.934 | 4.792 | 0.119 | 0.128 | 0.988 | $3.097 \pm 0.081$ | $3.001 \pm 0.087$ |
| | | ($\pm 0.060, \pm 0.107, \pm 0.084$) | $\pm 0.101$ | | | | | |
| | 5 | 7.794 | 7.759 | 0.144 | 0.145 | 1.021 | $5.030 \pm 0.097$ | $5.007 \pm 0.098$ |
| | | ($\pm 0.076, \pm 0.152, \pm 0.102$) | $\pm 0.139$ | | | | | |
| 50 | 1 | 2.299 | 1.845 | 0.117 | 0.164 | 0.994 | $1.317 \pm 0.079$ | $1.010 \pm 0.111$ |
| | | ($\pm 0.030, \pm 0.052, \pm 0.088$) | $\pm 0.078$ | | | | | |
| | 3 | 4.945 | 4.805 | 0.140 | 0.146 | 0.978 | $3.105 \pm 0.095$ | $3.010 \pm 0.098$ |
| | | ($\pm 0.049, \pm 0.103, \pm 0.103$) | $\pm 0.127$ | | | | | |
| | 5 | 7.799 | 7.764 | 0.187 | 0.188 | 1.022 | $5.033 \pm 0.127$ | $5.010 \pm 0.127$ |
| | | ($\pm 0.062, \pm 0.151, \pm 0.125$) | $\pm 0.173$ | | | | | |

Table 7.1 Scenario A (top) and B (bottom) mean dose estimates and standard deviations based on 100 simulation runs. The brackets in $\hat{\mu}_{CP}$ column read (PSE, QPSE, SE($\hat{\mu}_{CP}$)) and the reported $\hat{\alpha}$ values are an average of the ZINB-1 MLE.

| True Fraction $F$ (%) | True Dose $D$ (Gy) | $\hat{F}_{CP}$ | $\hat{F}_{CNB}$ | SD($\hat{F}_{CP}$) | SD($\hat{F}_{CNB}$) |
|---|---|---|---|---|---|
| 75 | 1 | 0.753 ±0.020 | 0.760 ±0.024 | 0.014 | 0.018 |
| | 3 | 0.750 ±0.014 | 0.750 ±0.014 | 0.002 | 0.003 |
| | 5 | 0.750 ±0.014 | 0.750 ±0.014 | 0.001 | 0.001 |
| 50 | 1 | 0.502 ±0.020 | 0.507 ±0.023 | 0.012 | 0.014 |
| | 3 | 0.500 ±0.016 | 0.501 ±0.016 | 0.002 | 0.002 |
| | 5 | 0.500 ±0.016 | 0.500 ±0.016 | 0.001 | 0.001 |
| 75 | 1 | 0.603 ±0.018 | 0.754 ±0.020 | 0.015 | 0.046 |
| | 3 | 0.728 ±0.014 | 0.749 ±0.015 | 0.005 | 0.007 |
| | 5 | 0.746 ±0.014 | 0.750 ±0.014 | 0.002 | 0.002 |
| 50 | 1 | 0.403 ±0.017 | 0.504 ±0.040 | 0.012 | 0.037 |
| | 3 | 0.485 ±0.016 | 0.500 ±0.017 | 0.004 | 0.005 |
| | 5 | 0.498 ±0.016 | 0.500 ±0.016 | 0.002 | 0.002 |

Table 7.2 Scenario A (top) and B (bottom) mean fraction estimates and standard deviations based on 100 simulation runs.

dose level and fraction used to generate this data is unknown. The whole process was repeated 100 times. Average-based estimates following the steps in Section 7.1.3 are displayed in Tables 7.1 and 7.2.

For *Scenario A*, we notice from Table 7.1 (top) that the CP and CNB appear to produce similar dose estimates which are almost identical to the true absorbed dose. Similarly, the fraction estimates from Table 7.2 (top) seem to be in agreement. By comparison, one could argue that the CP provides slightly better results, however the CNB estimates can be viewed as the corrected estimates after adjusting for the slight overdispersion (or underdispersion in the 3Gy sample). We note that each simulation explicitly produces an estimate for $\hat{\alpha}$, the mean of these values is stated in Table 7.1. For Poisson data it is expected that there should be no significant differences between the two methods.

The impact of the inclusion of extra zeros (or cells containing zero counts) for each dose in the whole-body sample (100% exposure data) can be immediately observed by

| $F$ (%) | $D$ (Gy) | *Merkle | MBD/Delta | *IAEA Simplified | MC (10000 trials) |
|---|---|---|---|---|---|
| 75 | 1 | (0.462, 2.016) | (0.500, 1.482) | (0.936, 1.048) | (0.489, 1.503) |
| | | (*0.453, 1.998*) | (*0.490, 1.470*) | (*0.925, 1.037*) | (*0.481, 1.490*) |
| | 3 | (1.970, 5.512) | (1.913, 4.109) | (2.921, 3.103) | (1.845, 4.321) |
| | | (*1.970, 5.511*) | (*1.903, 4.099*) | (*2.919, 3.103*) | (*1.843, 4.323*) |
| | 5 | (3.373, 9.022) | (3.240, 6.762) | (4.886, 5.119) | (3.134, 7.161) |
| | | (*3.373, 9.022*) | (*3.240, 6.762*) | (*4.886, 5.118*) | (*3.135, 7.160*) |
| | 1 | (0.715, 2.545) | (0.732, 1.880) | (1.243, 1.369) | (0.714, 1.927) |
| | | (*0.470, 2.032*) | (*0.507, 1.495*) | (*0.946, 1.058*) | (*0.500, 1.513*) |
| | 3 | (2.032, 5.664) | (1.971, 4.223) | (3.006, 3.191) | (1.905, 4.443) |
| | | (*1.963, 5.496*) | (*1.904, 4.098*) | (*2.911, 3.094*) | (*1.841, 4.309*) |
| | 5 | (3.393, 9.072) | (3.259, 6.801) | (4.914, 5.147) | (3.116, 7.229) |
| | | (*3.376, 9.030*) | (*3.240, 6.774*) | (*4.891, 5.123*) | (*3.093, 7.203*) |

Table 7.3 75% partially-irradiated sample dose estimation uncertainties from the Poisson (top) and NB1 (bottom) simulation for the CP (first row) and CNB (second row, *italic*), expressed in the form of 95% confidence intervals.

the decrease in fraction estimates. As expected, these only slightly deviate from the true value. We note that the dose estimates from each method remain constant and therefore are independent of zero-inflation. This is simply due to the number of zeros manually incorporated in the total number of observations, hence the value stays the same. The same holds true for the total focus count, in which only zeros are added.

The results from *Scenario B* are displayed in Table 7.1 (bottom). It is suggested, as a generalised criterion, that a dose estimate within 30% of the "true" localised dose is sufficiently accurate for radiological protection purposes [1]. This claim appears to be valid for all dose levels except for the 1Gy dose estimated from the CP method, lying slightly outside the interval. On the other hand, it is clear (also from the fraction estimates in Table 7.2 (bottom)) that the CNB method has accounted for both the base dispersion and additional dispersion as a result of zero-inflation, noting the difference in estimates between both methods in comparison to those from the simulated Poisson data. Seemingly, this appears to be most apparent for smaller doses, in our case the 1Gy samples.

As an extension, Table 7.3 details the 95% confidence limits for the dose estimates obtained from the simulated 75% exposed samples. It is evident that the uncertainties produced from the MBD and MC methods are of similar magnitude, while Merkle's method naturally provides larger upper limits. The results from the IAEA simplified approach are in accordance with the standard deviations reported in Table 7.1. Both the simulation and IAEA simplified do not consider the calibration curve standard errors, subsequently providing a narrow confidence interval. Initially, it would seem here that the simplified approach is most favourable, however we see from both the

overdispersed 1Gy and 3Gy CP dose intervals its drawbacks when dealing with a poor estimate.

A major limitation of our simulation setup is the exclusion of biological effects. In practice, the expected amount of "structural zeros" is likely to be smaller (i.e. less foci-free cells observed than expected), especially for lower exposures prone to background noise. Biological problems which can often arise from higher doses are also not captured by the simulation. To better understand the drawbacks of the simulation, in the following chapter we make the attempt to utilise the CP and CNB methods for purpose with practical H2AX and dicentric data.

# Chapter 8

# Application of CP and CNB methods

Following a major radiation accident, the estimated dose and fraction are crucial for the segregation of exposed patients into triage categories. Typically, these categories will consist of low dose (less than 1Gy), moderate dose between 1Gy and 3Gy and high doses above 3Gy. Despite the biological drawbacks, the results from the simulation in the previous Chapter still motivates the use of the contaminated negative binomial method to provide estimates accurate enough and thus confirm the triage placement of patients, in cases of an overdispersed aberration distribution. In this chapter, we will see that problems relating to the background level can contribute to poor estimates obtained from the CP and CNB methods. Adjustments which attempt to resolve these issues are discussed and compared.

## 8.1 Practical data analysis

We recall from Chapter 4 that the PHE-Foci2 Poisson calibration curve was found to be:

$$\lambda = 0.197(\pm 0.014) + 3.725(\pm 0.029)D, \tag{8.1.1}$$

where the intercept value of 0.197 represents the background level of foci. Table 8.1 provides both the dose and fraction estimates as compared to their physical quantities for each sample. Upon inspection, it is clear that the CP method was able to produce estimates which are closer to the true values, in respect, contradicting the results from our simulation. Furthermore, for the partially-exposed samples, it appears here that $\hat{\alpha}$ has made a insignificant contribution to the dose and fraction estimates from the CNB

| $F$ (%) | $D$ (Gy) | $\hat{\mu}_{CP}$ | $\hat{\mu}_{CNB}$ | $\hat{D}_{CP}$ | $\hat{D}_{CNB}$ | $\hat{F}_{CP}$ | $\hat{F}_{CNB}$ | $\hat{\alpha}$ |
|---|---|---|---|---|---|---|---|---|
| 30 | 0.75 | 2.725 | 2.561 | 0.679 | 0.635 | 0.371 | 0.395 | 0.462 |
| | | $(\pm 0.032, \pm 0.055, \pm 0.095)$ | $\pm 0.090$ | | | $\pm 0.016$ | $\pm 0.019$ | |
| | 1.5 | 4.738 | 4.364 | 1.219 | 1.119 | 0.367 | 0.399 | 1.907 |
| | | $(\pm 0.042, \pm 0.093, \pm 0.116)$ | $\pm 0.152$ | | | $\pm 0.015$ | $\pm 0.018$ | |
| | 3 | 8.692 | 8.509 | 2.280 | 2.232 | 0.352 | 0.360 | 3.131 |
| | | $(\pm 0.055, \pm 0.158, \pm 0.157)$ | $\pm 0.281$ | | | $\pm 0.015$ | $\pm 0.016$ | |
| 60 | 0.75 | 2.689 | 2.617 | 0.669 | 0.650 | 0.625 | 0.643 | 0.208 |
| | | $(\pm 0.041, \pm 0.060, \pm 0.073)$ | $\pm 0.068$ | | | $\pm 0.017$ | $\pm 0.019$ | |
| | 1.5 | 5.626 | 5.593 | 1.458 | 1.449 | 0.653 | 0.657 | 0.432 |
| | | $(\pm 0.061, \pm 0.109, \pm 0.094)$ | $\pm 0.107$ | | | $\pm 0.015$ | $\pm 0.015$ | |
| | 3 | 9.596 | 9.591 | 2.523 | 2.522 | 0.679 | 0.679 | 0.591 |
| | | $(\pm 0.081, \pm 0.171, \pm 0.119)$ | $\pm 0.149$ | | | $\pm 0.015$ | $\pm 0.015$ | |
| 100 | 0.75 | 3.090 | 3.090 | 0.777 | 0.777 | 0.975 | 0.977 | $< 10^{-4}$ |
| | | $(\pm 0.055, \pm 0.056, \pm 0.061)$ | $\pm 0.059$ | | | $\pm 0.009$ | $\pm 0.009$ | |
| | 1.5 | 5.923 | 5.919 | 1.537 | 1.536 | 0.975 | 0.975 | 0.085 |
| | | $(\pm 0.076, \pm 0.085, \pm 0.079)$ | $\pm 0.081$ | | | $\pm 0.005$ | $\pm 0.005$ | |
| | 3 | 11.440 | 11.440 | 3.019 | 3.019 | 0.993 | 0.993 | $< 10^{-4}$ |
| | | $(\pm 0.106, \pm 0.090, \pm 0.107)$ | $\pm 0.108$ | | | $\pm 0.003$ | $\pm 0.003$ | |

Table 8.1 Dose and fraction estimates corresponding to 30%, 60% and full-exposure conditions. The brackets in $\hat{\mu}_{CP}$ column read (PSE, QPSE, SE($\hat{\mu}_{CP}$)).

| $F$ (%) | $D$ (Gy) | MBD/Delta | *Merkle | *IAEA Simplified | MC (10000 trials) |
|---|---|---|---|---|---|
| 30 | 0.75 | (0.626, 0.731) | (0.634, 0.725) | (0.652, 0.706) | (0.627, 0.730) |
| | | (*0.585, 0.685*) | (*0.591, 0.680*) | (*0.609, 0.662*) | (*0.586, 0.683*) |
| | 1.5 | (1.155, 1.283) | (1.161, 1.279) | (1.183, 1.256) | (1.155, 1.284) |
| | | (*1.037, 1.201*) | (*1.063, 1.177*) | (*1.084, 1.154*) | (*1.036, 1.201*) |
| | 3 | (2.194, 2.366) | (2.202, 2.361) | (2.232, 2.330) | (2.191, 2.370) |
| | | (*2.082, 2.382*) | (*2.154, 2.311*) | (*2.184, 2.280*) | (*2.080, 2.383*) |
| 60 | 0.75 | (0.627, 0.711) | (0.624, 0.715) | (0.642, 0.697) | (0.629, 0.709) |
| | | (*0.611, 0.689*) | (*0.606, 0.695*) | (*0.623, 0.677*) | (*0.612, 0.688*) |
| | 1.5 | (1.405, 1.511) | (1.394, 1.523) | (1.419, 1.497) | (1.403, 1.512) |
| | | (*1.390, 1.508*) | (*1.385, 1.514*) | (*1.410, 1.488*) | (*1.388, 1.510*) |
| | 3 | (2.456, 2.590) | (2.441, 2.607) | (2.472, 2.575) | (2.455, 2.592) |
| | | (*2.440, 2.604*) | (*2.440, 2.606*) | (*2.471, 2.574*) | (*2.439, 2.606*) |
| 100 | 0.75 | (0.741, 0.813) | (0.729, 0.826) | (0.748, 0.806) | (0.744, 0.810) |
| | | (*0.742, 0.812*) | (*0.729, 0.826*) | (*0.748, 0.806*) | (*0.744, 0.810*) |
| | 1.5 | (1.491, 1.583) | (1.472, 1.604) | (1.497, 1.578) | (1.488, 1.587) |
| | | (*1.489, 1.583*) | (*1.471, 1.603*) | (*1.496, 1.577*) | (*1.486, 1.587*) |
| | 3 | (2.957, 3.081) | (2.928, 3.110) | (2.963, 3.075) | (2.952, 3.083) |
| | | (*2.956, 3.082*) | (*2.928, 3.110*) | (*2.963, 3.075*) | (*2.952, 3.083*) |

Table 8.2 95% confidence limits for the dose estimates obtained from CP (first row) and CNB (second row, *italic*) as reported in Table 8.1.

method. We deduce from the simulation and both Tables 8.1 and A.1 (Appendix A.8 - estimates for the PHE-Foci1 samples) that always $\hat{D}_{CP} \geq \hat{D}_{CNB}$ and $\hat{F}_{CP} \leq \hat{F}_{CNB}$. Being most apparent for the 30% samples, the attempt by the CNB method to account for any additional overdispersion can be viewed as having a reverse effect in the case when $F \leq \hat{F}_{CP} \leq \hat{F}_{CNB}$. After considering the uncertainty, we observe from Table 8.2 that only the 1.5Gy/60% (IAEA simplified excluded) and 100% samples yield confidence intervals which encompass the true dose. Nevertheless, $\hat{D}_{CP}$ and $\hat{D}_{CNB}$ appear to be within the suggested 30% margin of $D$.

By contrast, it can be observed from the dicentric results displayed in Table 8.3 that the fraction estimates have improved using the CNB in cases where $\hat{\alpha}$ is large (noting here that all $\hat{F}_{CP} < F$). For some samples, most notably the 50%/0.7Gy, one has to subsequently settle for worse estimates of $\hat{D}_{CNB}$. It should be highlighted that the doses used for the calibration curve range between 0.1-1Gy, with doses less than 1Gy considered very small. Estimation of such low doses in PBI scenarios is a subject which continues to receive ongoing attention [8]. It is plausible, in addition to only 3 dose levels (of wide range) used for construction of the calibration curve, that this could also explain some of the ambiguity in estimates from the 0.75Gy samples of the foci datasets.

| $F$ (%) | $D$ (Gy) | $\hat{D}_{CP}$ | $\hat{D}_{CNB}$ | $\hat{F}_{CP}$ | $\hat{F}_{CNB}$ | $\hat{\alpha}$ |
|---|---|---|---|---|---|---|
| 50 | 0.5 | 0.755 | 0.755 | 0.265 | 0.266 | 0.043 |
| | 0.7 | 0.795 | 0.519 | 0.306 | 0.537 | 0.369 |
| | 1 | 1.023 | 0.918 | 0.321 | 0.375 | 0.198 |
| 75 | 0.5 | 0.800 | 0.707 | 0.420 | 0.496 | 0.134 |
| | 0.7 | 0.822 | 0.821 | 0.473 | 0.474 | 0.002 |
| | 1 | 0.925 | 0.850 | 0.600 | 0.676 | 0.128 |
| 100 | 0.5 | 0.788 | 0.757 | 0.592 | 0.623 | 0.046 |
| | 0.7 | 0.829 | 0.767 | 0.788 | 0.874 | 0.096 |
| | 1 | 1.070 | 1.024 | 0.892 | 0.948 | 0.093 |

Table 8.3 Dose and fraction estimates corresponding to 50%, 75% and full-exposure conditions for PHE-Dicentric dataset.

Certainly, it is evident that the implications made from the simulation setup do not smoothly carry over to practical biomarker data. The consequence of the CP method overestimating the fraction of exposure leads to the CNB method being unable to make any improvements. This is also reflected in the dose uncertainties, in which we are unable to see the advantages of the MBD and MC methods in partial-exposures. However, it remains somewhat reassuring from Table 8.1 that all $\hat{F}$ do not deviate too greatly from $F$, with all estimates contained within $F \pm 10\%$. Assuming the true fraction is unknown, as initially implied, then we still require alternative procedures to firstly identify the cause(s) of fraction overestimation from the data characteristics (i.e. problem with the background level or too many zeros) and secondly if the estimates can at all be rectified.

## 8.2 Anomalies

Before making any attempt to modify our samples and/or apply alternative models, there are a few considerations to reflect upon. Firstly, the calculation of our PHE-Foci2 curve involves using an underdispersed sample, in turn, presenting some complications with its validity. As a solution, one could replace this sample with the corresponding 3Gy PHE-Foci1 sample or by using its dispersion index (with the PHE-Foci1 curve) to generate a new sample via NB1($n^* = 1000$, $\lambda = 0.766 + (1.700 \times 3)$; $\phi = 1.023$) then re-fit the calibration curve.

We observed in Chapter 1 that the 1.5Gy/100% sample contains a cell of 37 foci which can be considered extreme when compared with the next highest recorded cell of 20 foci. Indeed, remaining as true as possible to the data is important but equally it could be detrimental if one fails to identify any potential obscurities/outliers and make suitable

| $F$ (%) | $D$ (Gy) | $\hat{\mu}_M$ | $\hat{D}_M$ | $\hat{F}_M$ | $\hat{D}_{CP_{RE}}$ | $\hat{D}_{CNB_{RE}}$ |
|---|---|---|---|---|---|---|
| 30 | 0.75 | 2.916 | 0.731 | 0.347 | 0.769 | 0.718 |
| | 1.5 | 4.780 | 1.232 | 0.364 | 1.390 | 1.274 |
| | 3 | 8.693 | 2.284 | 0.352 | 2.610 | 2.553 |
| 60 | 0.75 | 2.885 | 0.723 | 0.583 | 0.758 | 0.735 |
| | 1.5 | 5.647 | 1.465 | 0.651 | 1.664 | 1.653 |
| | 3 | 9.596 | 2.527 | 0.679 | 2.889 | 2.887 |
| 100 | 0.75 | 3.237 | 0.817 | 0.931 | 0.881 | 0.881 |
| | 1.5 | 5.939 | 1.544 | 0.972 | 1.747 | 1.747 |
| | 3 | 11.44 | 3.023 | 0.993 | 3.458 | 3.458 |

Table 8.4 Estimated dose and fraction based on the method of moments and using the re-fitted calibration curve (replacing 3Gy sample and the 37 scored foci cell omitted from 1.5Gy sample).

adjustments. Removing this cell from the sample yields $\lambda = 0.197 + 3.719D$ (a slightly smaller slope means dose estimates for the other samples would barely differ) with QP dispersion of $\hat{\phi} = 1.038$. Additionally, the dispersion index $\hat{\delta}$ significantly decreases from 1.261 to 1.100 and now $\hat{\alpha} < 10^{-4}$ from the ZINB1 model. This results in the same fraction estimates but moderately smaller dose estimates of $\hat{D}_{CP} = \hat{D}_{CNB} = 1.532$, as compared with Table 8.1. Additionally, replacing the 3Gy sample with the corresponding 3Gy PHE-Foci1 sample, the calibration curve becomes $\lambda = 0.234 + (3.241D)$ with $\hat{\phi} = 1.314$. The re-calculated dose estimates are stated in Table 8.4 (indicated by "$RE$" subscript). Certainly, we see improved estimates for the partially-irradiated samples (some exception for the 1.5Gy/60% sample) but now overestimation in the WBI samples. Alternatively we could have used a non-Poisson/QP fitted curve, however we found there was relatively small deviation in resulting dose estimates (as expected from the similar coefficients reported earlier in Table 4.3).

From Table 8.4, it is reasonable to suggest for PBI that $\hat{D}_M$ could serve as a lower bound for $D$ and an upper bound in the case of WBI, remaining within the 30% discrepancy interval. Conversely, $\hat{F}_M$ could be viewed as being an upper bound for $F$ such that if $\hat{F}_{CP} \geq \hat{F}_M$ then this is an indication of fraction overestimation in PBI.

In Chapter 4 (Table 4.7) we discovered some evidence of cell death for the 30% exposure samples under the ZINB1, suggesting the estimated CNB fractions should be updated accordingly. Assuming the value for $\hat{\gamma}_1$ is known, then replacing $D$ with $\hat{D}_{CNB}$ in expression (4.4.2) and substituting into (4.4.1) provides increased fractions of $\hat{F}_{CNB} = 0.422$, 0.448 and 0.456 respectively. Subsequently, accounting for cell death will result in worse estimates in the case of fraction overestimation, however our evidence from Chapter 4 suggests we can have trust in the estimates presented in Tables 8.1 and 8.3.

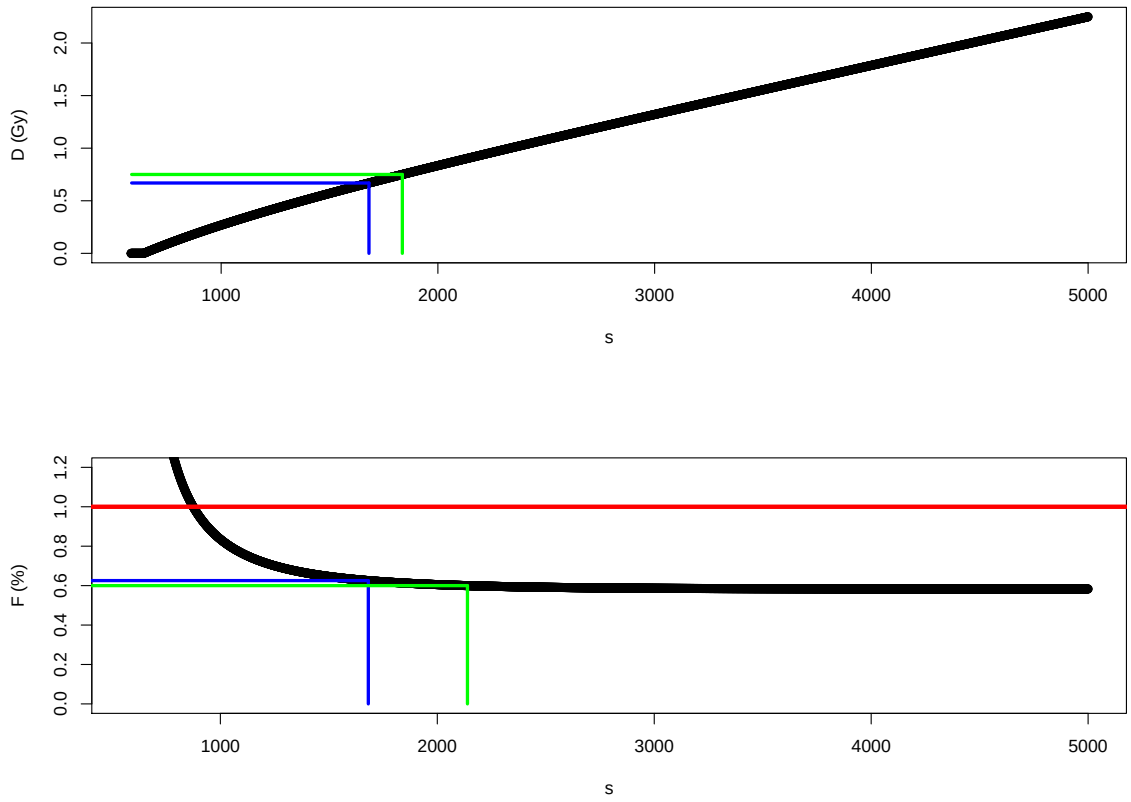



Fig. 8.1 Plots of $s^*$ against $\hat{D}_{CP}$ (top) and $\hat{F}_{CP}$ (bottom) for the 0.75Gy/60% sample. The blue line is used to indicate estimated values while the green line represents the true values. The scenario of both non-zero dose estimates and $F \leq 100\%$ (red line) is achieved when $s^* \geq 875$.

| $F$ (%) | $D$ (Gy) | min($s^*$) | $s^*$ when $\hat{F}_{CP} = F$ | Converging $\hat{F}_{CP}$ | $s^*$ when $\hat{D}_{CP} = D$ |
|---|---|---|---|---|---|
| | 0.75 | 427 | | 0.347 | 1093 |
| 30 | 1.5 | 453 | | 0.364 | 2113 |
| | 3 | 434 | | 0.352 | 4004 |
| | 0.75 | 875 | 2138 | | 1836 |
| 60 | 1.5 | 1053 | | 0.651 | 3778 |
| | 3 | 1137 | | 0.679 | 7721 |
| | 0.75 | 2673 | 2673 | | 2932 |
| 100 | 1.5 | 3576 | 3576 | | 5640 |
| | 3 | 4962 | 4962 | | 11292 |

Table 8.5 Values for $s^*$ for which both $\hat{D}_{CP} > 0$ and $\hat{F}_{CP} \leq 100\%$ (min($s$) column) and when true dose and fraction are obtained. The results in this table assume the same calibration curve (8.1.1) is used and that the observed zeros remains constant. Note that it is not possible to replicate this analysis for the CNB method since $\hat{\alpha}$ would require the individual foci frequencies (which total $s^*$).

It is clear from the expressions for the CP and CNB that the addition or subtraction of zeros does not alter $\hat{\mu}$ since the quantity $n^* - n_0^*$ remains unchanged. However, $\hat{\mu}$ is dependent on the total aberration count, $s^*$. From Table 8.5 we see that the 0.75Gy/60% sample is the only partially-irradiated sample in which both $D$ and $F$ can be reached by changing $s^*$, as shown in Figure 8.1, which alludes us to the fact that the problem with the remaining samples is primarily associated with the proportion (or lack) of zeros. With reference to the aforementioned 0.75Gy/60% sample, and recalling that always $\hat{D}_{CP} \geq \hat{D}_{CNB}$ and $\hat{F}_{CP} \leq \hat{F}_{CNB}$, at some value $s^* >> 2138$ (dependent on the breakdown of $s^*$ and hence the value of $\hat{\alpha}$) one eventually has both $\hat{D}_{CP} > \hat{D}_{CNB} > D$ and $\hat{F}_{CP} < \hat{F}_{CNB} \leq F$.

## 8.3 Adjusting for background exposure

### 8.3.1 Sensitivity of single-foci cells

As made evident from the values of $\hat{F}_M$, we identify in Figure 8.2 (overleaf) that only the 0.75Gy/60% sample contains the minimum number of foci-free cells to possibly distinguish between the structural zeros and sampling zeros. A feasible reason for observing less zeros is that low numbers of foci tends to trick the imaging software into detecting spurious foci background noise, by falsely enhancing the signals of minor granularities in the image which do not constitute foci. As a consequence there is a possibility that cells with reported small foci frequencies may have been under- or over-scored, depending on how experienced the scorer is. However, this does not directly

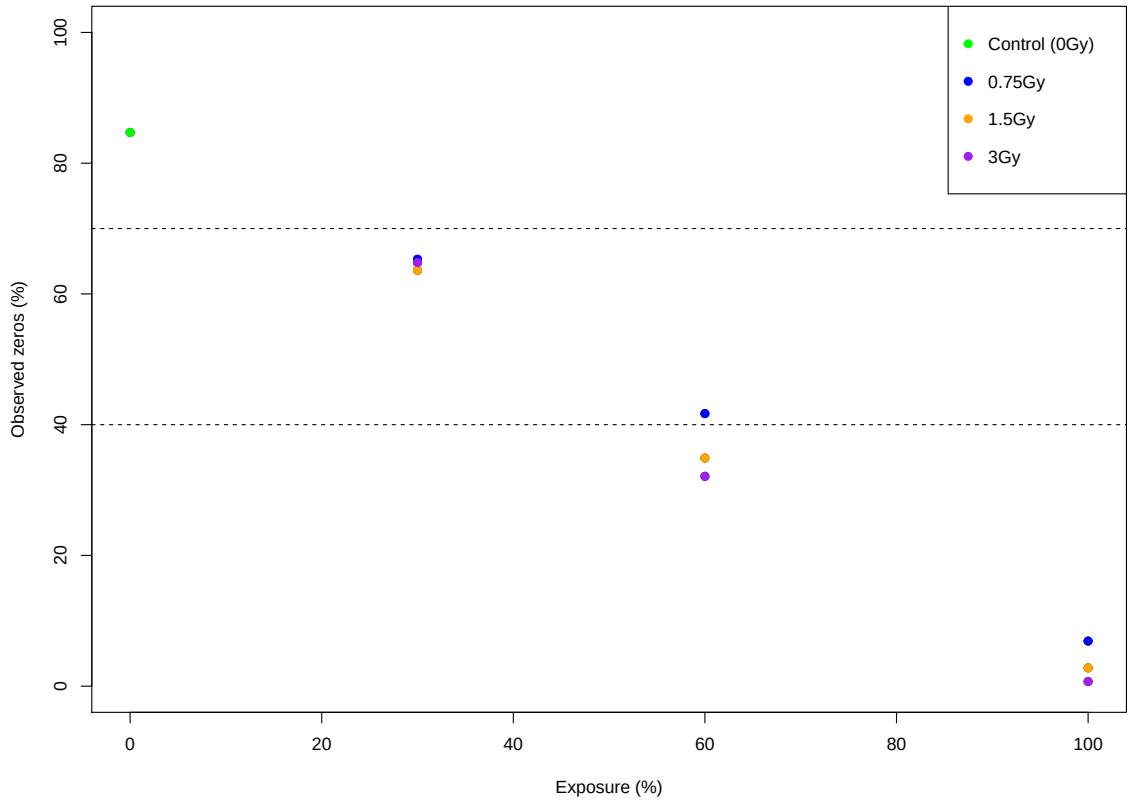



Fig. 8.2 Observed vs expected proportion of zeros for the PHE-Foci2 dataset. The dashed horizontal lines at 40% and 70% observed zeros indicate the structural zeros expected from the 60% and 30% partially-exposed samples respectively.

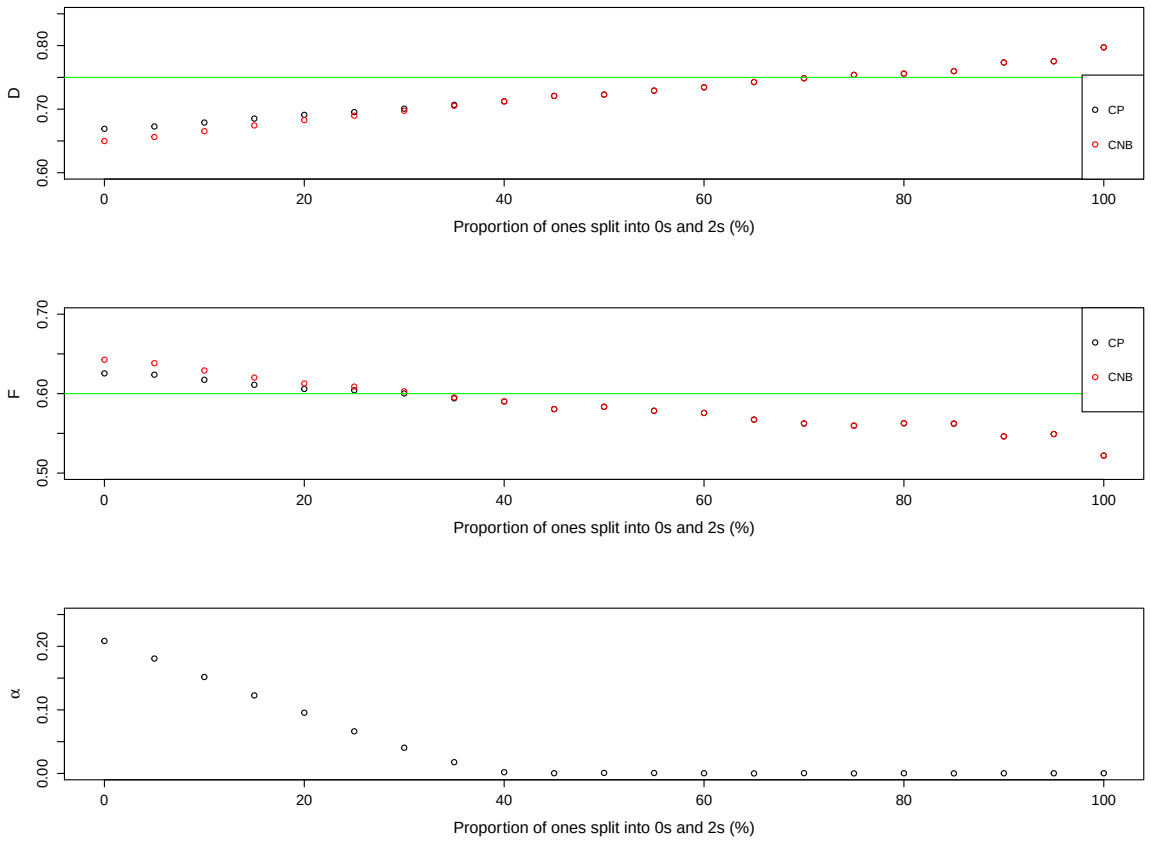



Fig. 8.3 Trend of dose (top), fraction (middle) and $\alpha$ (bottom) estimates over increasing 5% proportions of single-foci cells divided into 0s and 2s for the 0.75Gy/60% sample.

answer the question as to why the CP method remains favourable in the case of the 0.75Gy/60% sample.

We investigate this particular sample further by means of '*data sharpening*'. We choose to amend by splitting a proportion (in gradual intervals of 5%) of single-foci cells into zeros and twos with probabilities $q$ and $1-q$. To reflect the scenario that a single-foci cell is equally likely to be under- or over-scored and to keep the total numbers of foci in the modified samples to be roughly the same as in the original sample, we set $q = 0.5$. From a Bayesian perspective, $q = 0.5$ is equivalent to assuming no prior information is known relating to the scoring process. The mean estimates (based on 100 runs) for $\hat{D}$, $\hat{F}$ and $\hat{\alpha}$ are displayed above in Figure 8.3. We first notice that $\hat{D}_{CP}$ and $\hat{D}_{CNB}$ become identical when $\geq 35\%$ single-foci cells are split and the true dose is achieved at $\approx 70\%$ split. Similarly, $\hat{F}_{CP}$ and $\hat{F}_{CNB}$ share similar quantities for $\geq 30\%$ split. As we increase the proportion of ones split, the behaviour of $\hat{\alpha}$ is not monotonic decreasing but tends to a miniscule value greater than 0 (here $10^{-5}$).

When carrying out the same analysis on the remaining 30% and 60% exposed samples, we observed that $\hat{D}_{CNB}$ were either closer to the true dose $D$ or identical to $\hat{D}_{CP}$ only when $\hat{D}_{CP} \geq \hat{D}_{CNB} > D$. In contrast, it holds true that the $\hat{F}_{CNB}$

| $F$ (%) | $D$ (Gy) | $\hat{D}_{CP}$ | $\hat{D}_{CNB}$ | $\hat{F}_{CP}$ | $\hat{F}_{CNB}$ | $\hat{\alpha}$ |
|---|---|---|---|---|---|---|
| 30 | 0.75 | 0.622 | 0.561 | 0.324 | 0.357 | 0.574 |
| | 1.5 | 1.243 | 1.196 | 0.320 | 0.332 | 1.127 |
| | 3 | 2.334 | 2.322 | 0.322 | 0.324 | 1.637 |
| 60 | 0.75 | 0.635 | 0.613 | 0.580 | 0.599 | 0.226 |
| | 1.5 | 1.416 | 1.408 | 0.636 | 0.639 | 0.381 |
| | 3 | 2.513 | 2.513 | 0.661 | 0.661 | 0.341 |
| 100 | 0.75 | 0.744 | 0.744 | 0.949 | 0.949 | 0.001 |
| | 1.5 | 1.494 | 1.492 | 0.968 | 0.969 | 0.125 |
| | 3 | 2.965 | 2.965 | 0.993 | 0.995 | * |

Table 8.6 Dose and fraction estimates using a slope-only curve, $\lambda = 3.735D$.

were either closer to the true fraction $F$ or similar to $\hat{F}_{CP}$ when $\hat{F}_{CP} \leq \hat{F}_{CNB} < F$ (50%/0.7Gy does contradict this could be related to small exposure/dose problem?). This is simply due to the fact that the CNB method will implicitly produce $\hat{\mu}_{CNB}$ which are always smaller than $\hat{\mu}_{CP}$ for $\hat{\alpha} > 0$. We note that for some samples (particularly 3Gy dose), we were unable to deduce such conclusions (rather that the estimates from both the CP and CNB methods remain almost the same irrespective of the proportion split and values of $\hat{\alpha}$) due to the insufficient amount of single-foci cells to convert into zeros for the true dose and fraction to be reached.

### 8.3.2 Slope-only model

Ideally, instead of splitting zeros between cells of low counts, it is desirable to remove the background level of foci more randomly while maintaining roughly the original data structure. We subtract Poi(0.197) from each count then re-combine at random any produced negative counts with the positive non-zeros. Table 8.6 provides the dose and fraction estimates for each sample by means of such data sharpening process, using the calibration curve in (8.1.1) with intercept removed. By comparison with Table 8.1, all dose and fraction (exception of 100%/3Gy sample) estimates have now reduced and therefore improved in cases where either were overestimated.

### 8.3.3 Zero-and-one inflated Poisson

As opposed to attempting to remove the background level of foci in a sample, one could speculate that our problem resides with the small non-zero counts. In addition to zero, it is not uncommon for biomarker data to simultaneously contain inflated counts for an additional count value $r$>0. For such purposes, Lin and Tsai [65] proposed a zero-and-$r$ inflated Poisson (ZrIP) regression model. The ZrIP has pmf

$$P(Y_{ij}=y_{ij}|\mu_i,p_{i0},p_{i1},p_{i2})=\begin{cases}p_{i0}+p_{i2}e^{-\mu_i}, & \text{for } y_{ij}=0\\ p_{i1}+p_{i2}\frac{e^{-\mu_i}\mu_i^r}{r!}, & \text{for } y_{ij}=r\\ p_{i2}\frac{e^{-\mu_i}\mu_i^r}{r!}, & \text{when } y_{ij}\geq 1, y_{ij}\neq r\end{cases} \tag{8.3.1}$$

where $0\leq p_{il}\leq 1, l=0,1,2$ such that $\sum_{l=0}^{2}p_{il}=1$. The ZrIP contains the ZIP model when $p_{i1}=0$ and the Poisson model if $p_{i0}=p_{i1}=0$. A special case of the ZrIP model is the zero-and-one inflated Poisson model ($r=1$) with pmf defined by

$$P(Y_{ij}=y_{ij}|\mu_i,p_{i0},p_{i1})=\begin{cases}p_{i0}p_{i1}+(1-p_{i0})e^{-\mu_i}, & \text{if } y_{ij}=0,\\ p_{i0}(1-p_{i1})+(1-p_{i0})\mu_i e^{-\mu_i}, & \text{if } y_{ij}=1,\\ (1-p_{i0})e^{-\mu_i}\frac{\mu_i^{y_{ij}}}{y_{ij}!}, & \text{if } y_{ij}\geq 2.\end{cases} \tag{8.3.2}$$

We denote (8.3.2) as ZOIP $(p_{i0},p_{i1},\mu_i)$. When $p_{i0}=0$, the model is a Poisson model and for $p_{i1}=1$, returns the ZIP model. The ZOIP distribution was first used by Melkersson and Olsson [72] to analyse the annual number of dentist visits for a sample of adult Swedes. Zhang et al. [106] later studied the properties and likelihood-based inference methods of the ZOIP model.

Denoting $q_{i0}$ and $q_{i1}$ as the probability of the random variable $Y_{ij}$ being zero and one respectively, the pmf in (8.3.2) becomes

$$P(Y_{ij}=y_{ij}|\mu_i,q_{i0},q_{i1})=\begin{cases}q_{i0}, & \text{if } y_{ij}=0,\\ q_{i1}, & \text{if } y_{ij}=1,\\ \frac{(1-q_{i0}-q_{i1})}{1-e^{-\mu_i}-\mu_i e^{-\mu_i}}e^{-\mu_i}\frac{\mu_i^{y_{ij}}}{y_{ij}!}, & \text{if } y_{ij}\geq 2,\end{cases} \tag{8.3.3}$$

where $q_{i0}\geq 0, q_{i1}\geq 0$ and $q_{i0}+q_{i1}\leq 1$. The likelihood function of $(q_{i0},q_{i1},\mu_i)$ is then

$$L(q_{i0},q_{i1},\mu_i)\propto q_{i0}^{N_0}q_{i1}^{N_1}(1-q_{i0}-q_{i1})^{N-N_0-N_1}\frac{\mu_i^T}{(1-e^{-\mu_i}-\mu_i e^{-\mu_i})^{N-N_0-N_1}}e^{-(N-N_0-N_1)\mu_i}. \tag{8.3.4}$$

Here, $N_1=\#\{i:y_{ij}=1\}$ and $T=\sum_{y_{ij}\geq 2}y_{ij}$. Under the notation presented in (8.3.3), the mean and variance can be shown to be

$$E(y_{ij}|x_i)=q_{i1}+\frac{(1-q_{i0}-q_{i1})\mu_i}{1-e^{-\mu_i}-\mu_i e^{-\mu_i}}(1-e^{-\mu_i}), \tag{8.3.5}$$

$$\begin{aligned}\mathrm{Var}(y_{ij}|x_i) = q_{i1} + \frac{(1-q_{i0}-q_{i1})\mu_i}{1-e^{-\mu_i}-\mu_i e^{-\mu_i}}(1+\mu_i-e^{-\mu_i}) \\ - \left(q_{i1} + \frac{(1-q_{i0}-q_{i1})\mu_i}{1-e^{-\mu_i}-\mu_i e^{-\mu_i}}(1-e^{-\mu_i})\right)^2.\end{aligned} \tag{8.3.6}$$

It therefore follows that

$$\begin{aligned}\frac{\mathrm{Var}(y_{ij}|x_i)}{E(y_{ij}|x_i)} = \frac{\left(q_{i1} + \frac{(1-q_{i0}-q_{i1})\mu_i}{1-e^{-\mu_i}-\mu_i e^{-\mu_i}}(1+\mu_i-e^{-\mu_i})\right)}{\left(q_{i1} + \frac{(1-q_{i0}-q_{i1})\mu_i}{1-e^{-\mu_i}-\mu_i e^{-\mu_i}}(1-e^{-\mu_i})\right)} \\ - \left(q_{i1} + \frac{(1-q_{i0}-q_{i1})\mu_i}{1-e^{-\mu_i}-\mu_i e^{-\mu_i}}(1-e^{-\mu_i})\right).\end{aligned} \tag{8.3.7}$$

Based on the likelihood function in (8.3.4), the maximum likelihood estimates in the case of a single inhomogeneous sample for $q_0$ and $q_1$ are

$$\hat{q}_l = \frac{n_l}{n}, l = 0, 1 \tag{8.3.8}$$

and the MLE of $\mu$, $\hat{\mu}$, can be found by solving the following equation:

$$T(e^{\mu} - \mu - 1) - (n - n_0 - n_1)(e^{\mu} - 1)\mu = 0, \tag{8.3.9}$$

which can be solved numerically through optimisation/root-solver methods. (8.3.9) suggests that there is always a unique solution for $\mu_i$ if $n - n_0 - n_1 > 0$ (i.e. at least one count greater than 1). Given $\hat{\mu}$, one can obtain MLEs for $p_0$ and $p_1$ by using the transformation in (8.3.3), providing

$$\hat{p}_0 = \frac{\hat{q}_0 + \hat{q}_1 - (1+\hat{\mu})e^{-\hat{\mu}_i}}{1-(1+\hat{\mu})e^{-\hat{\mu}}}, \quad \hat{p}_1 = \frac{\hat{q}_0 - (1-\hat{p}_0)e^{-\hat{\mu}}}{\hat{p}_0}. \tag{8.3.10}$$

In our context, the ZOIP allows to determine the degree of inflation arising from both foci-free and single-foci scored cells. Although one could extend the ZOIP to allow the modelling of excess cells for a range of lower frequencies (say cells consisting of 0-3 foci) and to account for overdispersion not due to zero- and $r$-inflation (ZrINB), both of these methods would require computational implementation with support for the identity link function. The ZOIP is advantageous in the sense that estimates for $\hat{\mu}$, $\hat{p}_0$ and $\hat{p}_1$ (and hence $\hat{D}$ and $\hat{F}$) can easily be found using (8.3.9) and (8.3.10).

Comparison of Table 8.1 with Table 8.7 reveals a significant improvement in contrast to the CP method, albeit an exception of the 1.5Gy/60% sample in which a slightly worse dose estimate but a less deviating fraction estimate is observed. Interestingly, application of the ZOIP now results in both under- and over-estimation of the dose and fraction in the partially-irradiated samples. It can be seen from Table 8.7 that both the 1.5Gy samples provide $\hat{D} > D$ while $\hat{F} > F$ is identified only in the case

| $F$ (%) | $D$ (Gy) | $\hat{\mu}$ | $\hat{q}_0$ | $\hat{q}_1$ | $\hat{p}_0$ | $\hat{p}_1$ | $\hat{D}$ | $\hat{F}$ |
|---|---|---|---|---|---|---|---|---|
| 30 | 0.75 | 2.984 | 0.653 | 0.086 | 0.673 | 0.946 | 0.748 | 0.327 |
| | 1.5 | 5.926 | 0.636 | 0.090 | 0.721 | 0.881 | 1.538 | 0.279 |
| | 3 | 10.334 | 0.648 | 0.062 | 0.710 | 0.913 | 2.722 | 0.290 |
| 60 | 0.75 | 2.864 | 0.417 | 0.137 | 0.428 | 0.898 | 0.716 | 0.572 |
| | 1.5 | 6.152 | 0.349 | 0.073 | 0.413 | 0.842 | 1.599 | 0.587 |
| | 3 | 10.232 | 0.321 | 0.047 | 0.368 | 0.873 | 2.694 | 0.632 |
| 100 | 0.75 | 3.085 | 0.069 | 0.136 | 0.022 | | 0.775 | 0.978 |
| | 1.5 | 5.960 | 0.028 | 0.022 | 0.033 | 0.782 | 1.547 | 0.967 |
| | 3 | 11.449 | 0.007 | 0.001 | 0.008 | 0.889 | 3.021 | 0.992 |

Table 8.7 Results from the ZOIP. Dose estimates are based on using the Poisson calibration curve (8.1.1) and reported $\hat{F}$ values are calculated via $1 - \hat{p}_0$. A value of $\hat{p}_1 > 1$ was obtained for the 100/0.75Gy sample (a consequence of $\hat{q}_1 > \hat{q}_0$) and was therefore omitted.

of the 0.75Gy/30% and 3Gy/60% samples. The increased values for $\hat{D}$ is simply a consequence of a larger $\hat{\mu}$ after accounting for the proportion of ones. Assuming no background irradiation, it is expected that $\hat{q}_0 \to 1$ as $F \to 0$. It is clear there is some discrepancy in the $\hat{q}_1$ between the 0.75Gy/60% and 0.75Gy/100% sample. We note that the latter is the only sample in which $\hat{q}_1 \geq \hat{q}_0$. Although we anticipate that the proportion of foci-free cells should be relatively small, it is logical that $\hat{q}_1 \geq \hat{q}_0$ should hold in any WBI scenario.

# Chapter 9

# Discussion

The gamma-H2AX assay as a biomarker for DSBs and radiation exposure has been firmly established in the literature. Compared to the "gold-standard" dicentric assay, where dose estimates can only be obtained in a costly and time-intensive process (chromosomes need to be cultured to metaphase which takes 2-3 days), the gamma-H2AX assay has the advantage that foci appear at the sites of DSBs within minutes which would allow for much quicker triage in the case of a large-scale radiation incident. However, only a relatively limited amount of work has been carried out towards enhancing the analysis methods in order to improve the accuracy and reliability of this assay.

Statistical concepts developed to deal with detecting partial body exposure in the context of the dicentric assay will not immediately carry over to protein-based assays as one will observe overdispersion irrespective of the exposure pattern in the latter. As an exception to this, we introduced in Chapter 1 the PHE-Dicentric dataset which also exhibited overdispersion in both partial and whole body irradiated samples. This means there is a need to consider how the dispersion index (or model dispersion) can be used to support estimates of partial body exposure, with emphasis on the gamma-H2AX assay. In addition, part of this thesis was to determine the nature of the overdispersion, that is, separating the contribution due to zero-inflation, and the consequences on dose (and fraction) estimates when using the standard methodology.

In Chapter 2 we discussed the uncertainty estimation techniques traditionally used in conjunction with whole-body exposures, in which the aberration distribution is assumed to be Poisson. Given present overdispersion in both PBI and WBI, as verified both through the dispersion indices in Chapter 3 and model dispersions in Chapter 4, these methods become invalid and no longer applicable (particularly for the H2AX assay).

In Chapter 4 we reviewed some of the alternative count data regression models to the Poisson which can be used to handle overdispersion and/or zero-inflation. Poisson

estimates are consistent when the variance is proportional (not just equal) to the mean. Therefore, and as we have shown, Poisson standard errors tend to be conservative in the presence of overdispersion. Choosing to disregard overdispersion in the analysis leads to underestimation of the calibration curve coefficient standard errors, and consequent overstatement of significance in hypothesis testing. If the Poisson is the chosen model and if we are sure that the lack of fit is not due to poor specification of the systematic part of the model, the standard errors need to be corrected by employing a quasi-likelihood approach. Another possibility is to utilise the negative binomial models (NB1 and NB2), which are based on maximum likelihood methods. A comparison of the intercept and slope parameter estimates along with their standard errors showed that there were some slight notable differences between the two approaches. A primary advantage of utilising a negative binomial regression is that it does not assume homogeneous variances but instead models heterogeneity in variance via the dispersion parameter $\alpha$.

Upon computation of the QP (and NB1) dispersions for both H2AX PHE-Foci datasets, we identified a significance difference in their realised quantities. Based on knowledge from the literature, a reported $\hat{\phi} > 1$ for H2AX whole-body calibration data would be viewed as being correct, however this could easily be misinterpreted since one may assume present overdispersion across all samples used in the fitting of the calibration curve (as is the case for the PHE-Foci1 dataset). We also note that our data consists of only 3 dose points (excluding 0Gy samples) - a more accurate curve (and hence dispersion estimate) requires an extended range of doses, ideally with multiple samples per dose. For this reason, if laboratories were to provide estimates of $\hat{\phi}$ and $\hat{\alpha}$ then it should be necessary to report not only the doses but also the summary statistics of those samples used.

The relationship between dispersion and dose is a subject which requires further attention. While for PBI we have seen from Chapter 3 that dispersion increases with dose (and decreases with exposure for exposure levels $> 30\%$), it appears more complex to determine the exact behaviour in the case of WBI. For zero-inflation modelled as constant, the non-constant dispersion pursues a positive linear association with level of dose. Meanwhile, when the zero-inflation parameter is variable, dispersion appears to peak at some dose less than 1Gy before decreasing with dose, eventually becoming equivalent and smaller than the constant QP/NB1 dispersion. Depending on the type of zero-inflated model being used, the maximum dispersion could potentially serve as a cutoff point for classification of a small and large dose. Certainly, this remains an area for further investigation, however it is natural to consider separate calibration curves for low and larger dose estimation. Although additionally in Chapter 4 we observed some evidence for cell death in the PHE-Foci2 30% exposure data using the ZINB1, the resulting dose estimates did not greatly change - generally cell death can be assumed negligible for the H2AX assay.

In Chapter 5 we discussed further how the dispersion parameter can be used as information against the Poisson model through the use of the likelihood ratio and Wald test. Since it can be shown that the negative binomial regression is a special case of the Poisson when $\alpha = 0$, this is equivalent to testing the null hypothesis $H_0 : \alpha = 0$ against the alternative $H_1 : \alpha > 0$. As expected, both tests indicated sufficient evidence against equidispersion and hence a Poisson fit. Additionally, we carried out tests for zero-inflation, that is, the null hypothesis $H_0 : p = 0$ against the alternative $H_1 : p > 0$, which supported the results from Chapter 2. We then proceeded to compare these tests with score tests proposed by Dean and Lawless which have the advantage over the likelihood ratio and Wald test in that it only requires the parameter estimated under the null hypothesis. A consequence of its simplicity meant the test statistics associated with the score test were found to be much larger in comparison to the likelihood ratio and Wald test. Despite this, all three tests were in favour of the NB and zero-inflated models over the Poisson. As an extension to strengthen the analysis in this chapter, a simulation study could be used to examine the properties of each test, in terms of nominal level attainment and statistical power, whereby a higher proportion of rejections of the null hypothesis at a given significance level would signify better power for that test.

In Chapter 6, we conducted analysis on the BfS-Foci dataset which, in comparison to the PHE H2AX datasets, consists of multiple experiments per slide. When applying a quasi-Poisson regression to both the raw foci distribution and aggregated counts, we discovered a very interesting phenomena. The aggregated dispersion was found to be $\hat{\phi}_{agg} = 147.99$ but this estimate is much larger than the "correct" raw dispersion value of $\hat{\phi} = 1.22$. We attempted to uncover what is causing this mass inflation by investigating the behaviour of the aggregated dispersion when violating the independence assumption of foci counts. We distinguished that there are two effects which jointly impact on the estimated dispersion; an increased variance effect and a dependency effect. The increased variance effect was explained for both the parametric and non-parametric bootstrap simulation method by the larger dispersion variances observed after aggregation. The dependency effect describes any unobserved variance between cells for a given dose.

We further presented, for a single dose, three heterogeneity scenarios to explain the significance of this dependency effect on the aggregated dispersion estimates. In our context, there are many experimental factors in the scoring of cells which can contribute to unobserved variance and hence an increase in dispersion when using aggregated foci counts. Wherever possible, the consensus appears to be that one should always try to use the full frequency distribution as lower dispersion values enable potential detection of partial-exposures. If we have only yields available (not raw counts), then of course one needs to work with a higher dispersion magnitude. It would be of interest in future work to explore these effects in more detail for other biomarkers such as the gene-expression and micronuclei assays.

Chapters 7 and 8 focus on the scenario in which a practitioner is provided with a random blood sample and the steps thereafter the practitioner needs to take in order to estimate the irradiated dose and fraction for which the patient has been exposed to. We note that part of the analysis involves fitting the ZIP or ZINB1 model meaning the scored raw counts must be used. The data type that is used for generating the Poisson calibration curve remains irrelevant, however, for reasons discussed in Chapter 6, the data type should be stated when working with uncertainty and hence the parameter standard errors and/or variance-covariance matrix (adjusted for overdispersion).

We illustrated successfully through simulated H2AX data the ability of the novel contaminated negative binomial to provide more accurate dose and fraction estimates, as compared to those resulting from the traditional contaminated Poisson. Upon application to real data, we identified that problems begin to materialise which relate to the background level. More specifically, it became clear that when the fraction has been overestimated then the attempt by the CNB to account for additional overdispersion will skew estimates in the wrong direction. In cases where the fraction had been underestimated (for example the 80% exposure PHE-Foci1 samples), we were unable to see any improvement by the CNB method (hence estimates remained the same as those from the CP) due to very small $\hat{\alpha}$. The results from the PHE-Dicentric samples were able to showcase better fraction estimates for larger values of $\hat{\alpha}$ but at the cost of worse dose estimates.

In Chapter 8, we investigated eliminating the background level via data sharpening techniques and employing a slope-only model. While these approaches did improve estimates, we believe our problem resides not only with the quantity of aberration-free or single-aberration cells but generally with the frequency distribution of lower counts. For this reason, we exercised the ZOIP model which was able to significantly amend dose and fraction estimates and, in the process, fix the issue of fraction overestimation in most samples. One could possibly extend the ZOIP to simultaneously account for larger inflated counts, say cells containing two and three aberrations, but extra care is required in separating natural to radiation-induced foci/dicentrics for larger counts.

# Appendix A

## A.1 Regression slope standard error

Once the fitted model has been implemented in R, the variance-covariance matrix for the quasi-Poisson response can be found by:

```
fit.li.poi1 <- glm(Foci~I(Dose) +  I(dTime24),
family="quasipoisson"(link="identity"), data=PHE1calidata1) #24h
#linear combined
coef(fit.li.poi1) # Parameter estimates

(Intercept) I(Dose)  I(dTime24)
  0.54141   1.83993   -0.92002

vcov(fit.li.poi1) # Covariance matrix

             (Intercept)       I(Dose)    I(dTime24)
(Intercept)  7.203263e-04 -0.0003910566  1.079257e-05
I(Dose)     -3.910566e-04  0.0036630097 -3.456569e-03
I(dTime24)   1.079257e-05 -0.0034565689  5.395024e-03
```

Extracting the relevant elements above, the standard error of the 24h linear combined slope can be computed through:

```
sqrt(c(1,1)%*%vcov(fit.li.poi1)[2:3,2:3]%*%c(1,1))
0.04631303
```

## A.2 Poisson and QP sampling error

Recall from Chapter 2 that the Poisson likelihood is defined by

$$L = \prod_{i,j} f(y_{ij}|x_i) = \prod_{i,j} e^{-\mu_i} \frac{\mu_i^{y_{ij}}}{y_{ij}!} \propto \prod_i e^{-n_i\mu_i} \mu_i^{\sum_j y_{ij}},$$

thus the log-likelihood is

$$\ell = \sum_{i=1}^{k} \sum_{j=1}^{n_i} [y_{ij} ln(\mu_i) - \mu_i - ln(y_i!)]$$

Differentiating with respect to $\mu_i$ and equating to zero yields the MLE

$$\begin{aligned} \frac{\partial \ell}{\partial \mu_i} &= \sum_{j=1}^{n_i} \left[ \frac{y_{ij}}{\mu} - 1 \right] = 0 \\ y_i n_i - \hat{\mu} n_i &= 0 \\ \hat{\mu} &= y_i, \end{aligned}$$

i.e. the slide-wise mean. This now allows to implement its standard error, namely

$$\mathrm{SE}(\hat{\mu}_i) = \sqrt{\hat{\mu}_i} = \sqrt{y_i}$$

and so

$$\mathrm{SE}\left( \sum_{j=1}^{n_i} y_{ij} \right) = \left( \sum_{j=1}^{n_i} \mathrm{SE}^2(y_{ij}) \right) = \sqrt{\sum_{j=1}^{n_i} \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}} = \sqrt{\sum_{j=1}^{n_i} y_{ij}}.$$

The variance associated with the foci/dicentric yield is then

$$\mathrm{Var}(y_i) = \mathrm{SE}^2 \left( \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \right) = \left( \frac{1}{\sqrt{n_i}} \sqrt{n_i} \right)^2 = \frac{y_i}{n_i},$$

therefore the sampling error remains

$$\mathrm{SE}(y_i) = \sqrt{\mathrm{Var}(y_i)} = \sqrt{\frac{y_i}{n_i}}.$$

However, this does ignore the overdispersion stemming from intra– and inter–individual variation. From the theory of the simple exponential family, under the presence of dispersion one has

$$\mathrm{Var}(\mu_i) = \phi \mathrm{SE}^2(\mu_i).$$

therefore the QPSE is defined by

$$\mathrm{QPSE}(y_i) = \sqrt{\frac{\hat{\phi}_{QP} y_i}{n_i}}$$

## A.3 ZIP and ZINB MLE with covariates

It is clear that being a finite mixture, the ZIP distribution is not a member of the exponential family distribution and so standard glm fitting procedures will not be adequate. To obtain the parameter estimates of ZIP regression models, $\boldsymbol{\beta}$ and $\hat{p}_i$, the Newton-Raphson method or the method of Fisher scoring can be used. Given the log-likelihood function in Section 4.3.1, the derivatives with respect to $\boldsymbol{\beta}$ and $p$ are

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_t} &= \frac{\partial \ell}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_t} \\ &= \sum_{j=1}^{n_i}\left(I_{(y_{ij}=0)}\left[-\frac{(1-p_i)e^{-\mu_i}}{p_i+(1-p_i)e^{-\mu_i}}\right] + I_{(y_{ij}>0)}\left[-1+\frac{y_{ij}}{\mu_i}\right]\right)\frac{\partial g(\mu_i)}{\partial \beta_t},\end{aligned}$$

for $t = 0, 1, ..., v$;

$$\frac{\partial \ell}{\partial p_i} = \sum_{j=1}^{n_i}\left(I_{(y_{ij}=0)}\left[\frac{1-e^{-\mu_i}}{p_i+(1-p_i)e^{-\mu_i}}\right] + I_{(y_{ij}>0)}\left[\frac{-1}{1-p_i}\right]\right);$$

and

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \beta_t \partial \beta_u} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Bigg(&I_{(y_{ij}=0)}\left[\frac{-e^{-\mu_i}\left[(1-\mu_i)p_i+(1-p_i)e^{-\mu_i}\right](1-p_i)\mu_i}{(p_i+(1-p_i)e^{-\mu_i})^2}\right] \\ &+ I_{(y_{ij}>0)}[-\mu_i]\Bigg)\frac{\partial^2 g(\mu_i)}{\partial \beta_t \partial \beta_u};\end{aligned}$$

$$\frac{\partial^2 \ell}{\partial p_i^2} = \sum_{j=1}^{n_i}\left(I_{(y_{ij}=0)}\left[\frac{-(1-e^{-\mu_i})^2}{(p_i+(1-p_i)e^{-\mu_i})^2}\right] + I_{(y_{ij}>0)}\left[\frac{-1}{(1-p_i)^2}\right]\right);$$

$$\frac{\partial^2 \ell}{\partial \beta_t \partial p_i} = \frac{\partial^2 \ell}{\partial p_i \partial \beta_t} = \sum_{j=1}^{n_i}\left(I_{(y_{ij}=0)}\left[\frac{\mu_i e^{-\mu_i}}{(p_i+(1-p_i)e^{-\mu_i})^2}\right]\right)\frac{\partial g(\mu_i)}{\partial \beta_t}.$$

Given that

$$\mathrm{E}\left[I_{(y_{ij}=0)}\right] = P(Y_{ij}=0) = p_i + (1-p_i)e^{-\mu_i}, \tag{A.3.1}$$

$$\mathrm{E}\left[I_{(y_{ij}>0)}\right] = P(Y_{ij}>0) = (1-p_i)(1-e^{-\mu_i}), \tag{A.3.2}$$

we have

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial\beta_t \partial\beta_u}\right] = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(\frac{e^{-\mu_i}\left[(1-\mu_i)p_i + (1-p_i)e^{-\mu_i}\right](1-p_i)\mu_i}{p_i + (1-p_i)e^{-\mu_i}} + \mu_i(1-p_i)(1-e^{-\mu_i})\Big)\frac{\partial^2 g(\mu_i)}{\partial\beta_t\partial\beta_u};$$

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial p_i^2}\right] = \sum_{j=1}^{n_i}\left(\frac{(1-e^{-\mu_i})^2}{p_i + (1-p_i)e^{-\mu_i}} + \frac{(1-e^{-\mu_i})}{(1-p_i)}\right);$$

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial\beta_t \partial p_i}\right] = \sum_{j=1}^{n_i}\left(\frac{-\mu_i e^{-\mu_i}}{p_i + (1-p_i)e^{-\mu_i}}\right)\frac{\partial g(\mu_i)}{\partial\beta_t}.$$

Hence the estimates of $\boldsymbol{\beta}$ and $p_i$ at the $(m+1)$th iteration, denoted by $\boldsymbol{\beta}^{(m+1)}$ and $p^{(m+1)}$, are given by

$$\begin{pmatrix}\boldsymbol{\beta}^{(m+1)}\\ p^{(m+1)}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta}^{(m)}\\ p^{(m)}\end{pmatrix} + \left[\mathrm{I}^{(m)}(\boldsymbol{\beta}, p)\right]^{-1}\mathrm{Sc}^{(m)}(\boldsymbol{\beta}, p),$$

where the score vector and expected information matrix are evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$ and $p = p^{(m)}$. With good starting values for $\boldsymbol{\beta}^{(0)}$, $p^{(0)}$, the iterative scheme converges in a few steps, with the asymtotic variance-covariance matrix for $(\hat{\boldsymbol{\beta}}, \hat{p})$ automatically provided in the final iteration.

For the ZINB regression, the first-order derivatives with respect to $\hat{\boldsymbol{\beta}}, p$ and $\alpha$ are

$$\begin{aligned}\frac{\partial \ell}{\partial \mu_i} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big( & I_{(y_{ij}=0)}\left[\frac{(1-p_i)(1+\alpha\mu_i^c)^{\frac{\mu_i^{1-c}}{\alpha}}\left(-\frac{c}{1+\alpha\mu_i^c} - \frac{(1-c)\mu_i^{-c}ln(1+\alpha\mu_i^c)}{\alpha}\right)}{p_i + (1-p_i)(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}}\right]\\
& + I_{(y_{ij}>0)}\Big[\frac{(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}}{(1-p_i)\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}\right)}\left(1+\frac{\mu_i^{-c}}{\alpha}\right)^{y_{ij}}\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}\right)\Big[(1-p_i)\\
& (1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(1+\frac{\mu_i^{-c}}{\alpha}\right)^{-y_{ij}}\left(-\frac{c}{1+\alpha\mu_i^c} - \frac{(1-c)\mu_i^{-c}ln(1+\alpha\mu_i^c)}{\alpha}\right)\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)\\
& - \frac{1}{\alpha}\Big[(1-c)(1-p_i)\mu_i^{-c}(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(1+\frac{\mu_i^{-c}}{\alpha}\right)^{-y_{ij}}\psi^{(0)}\left(\frac{\mu_i^{1-c}}{\alpha}\right)\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)\\
& + (1-c)(1-p_i)\mu_i^{-c}(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(1+\frac{\mu_i^{-c}}{\alpha}\right)^{-y_{ij}}\psi^{(0)}\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)\\
& + c(1-p_i)y_{ij}\mu_i^{-c-1}(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(1+\frac{\mu_i^{-c}}{\alpha}\right)^{-y_{ij}-1}\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)\Big]\Big]\Big]\Big);\end{aligned}$$

$$\frac{\partial \ell}{\partial \beta_t} = \frac{\partial \ell}{\partial \mu_i}\frac{\partial g(\mu_i)}{\partial \beta_t};$$

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Bigg( I_{(y_{ij}=0)}\left[\frac{(1-p_i)(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(-\frac{\mu_i}{\alpha(1+\alpha\mu_i^c)}+\frac{\mu_i^{1-c}ln(1+\alpha\mu_i^c)}{\alpha^2}\right)}{p_i+(1-p_i)(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}}\right] + I_{(y_{ij}>0)}\Bigg[$$

$$\frac{(1+\alpha\mu_i^c)^{\frac{\mu_i^{1-c}}{\alpha}}}{(1-p_i)\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}\right)}\left(1+\frac{\mu_i^{-c}}{\alpha}\right)^{y_{ij}}\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}\right)\Bigg[(1-p_i)$$

$$(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(1+\frac{\mu_i^{-c}}{\alpha}\right)^{-y_{ij}}\left(-\frac{\mu_i}{\alpha(1+\alpha\mu_i^c)}+\frac{\mu_i^{1-c}ln(1+\alpha\mu_i^c)}{\alpha^2}\right)\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)$$

$$+\frac{1}{\alpha^2}\Bigg[(1-p_i)\mu_i^{1-c}(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(1+\frac{\mu_i^{-c}}{\alpha}\right)^{-y_{ij}}\psi^{(0)}\left(\frac{\mu_i^{1-c}}{\alpha}\right)\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)$$

$$-(1-p_i)\mu_i^{1-c}(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(1+\frac{\mu_i^{-c}}{\alpha}\right)^{-y_{ij}}\psi^{(0)}\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij})\right)\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)$$

$$+(1-p_i)y_{ij}\mu_i^{-c}(1+\alpha\mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(1+\frac{\mu_i^{-c}}{\alpha}\right)^{-y_{ij}-1}\Gamma\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)\Bigg]\Bigg]\Bigg]\Bigg);$$

$$\frac{\partial \ell}{\partial p_i} = \sum_{j=1}^{n_i}\left( I_{(y_{ij}=0)}\left[\frac{1-(1+\alpha\mu_i^c)^{\frac{-\mu_i^{1-c}}{\alpha}}}{(1-p_i)(1+\alpha\mu_i^c)^{\frac{-\mu_i^{1-c}}{\alpha}}+p_i}\right] - I_{(y_{ij}>0)}\left[\frac{1}{1-p_i}\right]\right).$$

.

The second-order derivatives are

$$\frac{\partial^2 \ell}{\partial p_i^2} = -\sum_{j=1}^{n_i}\left( I_{(y_{ij}=0)}\left[\frac{(1-(1+\alpha\mu_i^c)^{\frac{-\mu_i^{1-c}}{\alpha}})^2}{((1-p_i)(1+\alpha\mu_i^c)^{\frac{-\mu_i^{1-c}}{\alpha}}+p_i)^2}\right] + I_{(y_{ij}>0)}\left[\frac{1}{(1-p_i)^2}\right]\right);$$

$$\frac{\partial^2 \ell}{\partial \mu_i^2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Bigg( I_{(y_{ij}=0)}\Bigg[ \frac{(1-p_i)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(\frac{\alpha c^2\mu_i^{c-1}}{\left(1+\alpha\mu_i^c\right)^2} - \frac{(1-c)c}{\mu_i\left(1+\alpha\mu_i^c\right)} + \frac{(1-c)c\mu_i^{-c-1}ln\left(1+\alpha\mu_i^c\right)}{\alpha}\right)}{(1-p_i)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}} + p_i}$$

$$- \frac{(1-p_i)^2\left(1+\alpha\mu_i^c\right)^{-\frac{2\mu_i^{1-c}}{\alpha}}\left(-\frac{c}{1+\alpha\mu_i^c} - \frac{(1-c)\mu_i^{-c}ln\left(1+\alpha\mu_i^c\right)}{\alpha}\right)^2}{\left((1-p_i)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}} + p_i\right)^2}$$

$$+ \frac{(1-p_i)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(-\frac{c}{1+\alpha\mu_i^c} - \frac{(1-c)\mu_i^{-c}ln\left(1+\alpha\mu_i^c\right)}{\alpha}\right)^2}{(1-p_i)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}} + p_i}\Bigg] + I_{(y_{ij}>0)}\Bigg[$$

$$\frac{\mu_i^{-2(c+1)}}{\alpha^2\left(1+\alpha\mu_i^c\right)^2}\Big[\Big(2\alpha^3c^2\mu_i^{3c+1} - \alpha^3c^2y_{ij}\mu_i^{3c} - \alpha^3c^2\mu_i^{3c+1}ln\left(1+\alpha\mu_i^c\right) - \alpha^3c\mu_i^{3c+1} - \alpha^3cy_{ij}\mu_i^{3c}$$

$$+ \alpha^3c\mu_i^{3c+1}ln\left(1+\alpha\mu_i^c\right) + \alpha^2c^2\mu_i^{2c+1} + \alpha^2c^2\mu_i^{2c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha} + y_{ij}\right) -$$

$$\alpha^2c^2\mu_i^{2c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right) - 2\alpha^2c^2\mu_i^{2c+1}ln\left(1+\alpha\mu_i^c\right) - \alpha^2c\mu_i^{2c+1} - \alpha^2cy_{ij}\mu_i^{2c} +$$

$$\alpha^2\mu_i^{2c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha} + y_{ij}\right) - 2\alpha^2c\mu_i^{2c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha} + y_{ij}\right) - \alpha^2\mu_i^{2c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right)$$

$$+ 2\alpha^2c\mu_i^{2c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right) + 2\alpha^2c\mu_i^{2c+1}ln\left(1+\alpha\mu_i^c\right) + 2\alpha c^2\mu_i^{c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha} + y_{ij}\right)$$

$$- 2\alpha c^2\mu_i^{c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right) + c^2\mu_i^2\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha} + y_{ij}\right) - c^2\mu_i^2\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right) -$$

$$\alpha c^2\mu_i^{c+1}ln\left(1+\alpha\mu_i^c\right) + 2\alpha\mu_i^{c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha} + y_{ij}\right) - 4\alpha c\mu_i^{c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha} + y_{ij}\right)$$

$$- 2\alpha\mu_i^{c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right) + 4\alpha c\mu_i^{c+2}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right) - 2c\mu_i^2\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha} + y_{ij}\right) +$$

$$\mu_i^2\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha} + y_{ij}\right) + 2c\mu_i^2\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right) - \mu_i^2\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right) +$$

$$\alpha c\mu_i^{c+1}ln\left(1+\alpha\mu_i^c\right) + \alpha\left(c-1\right)c\mu_i^{c+1}\left(1+\alpha\mu_i^c\right)^2\psi^{(0)}\left(\frac{\mu_i^{1-c}}{\alpha} + y_{ij}\right) -$$

$$\alpha\left(c-1\right)c\mu_i^{c+1}\left(1+\alpha\mu_i^c\right)^2\psi^{(0)}\left(\frac{\mu_i^{1-c}}{\alpha}\right)\Big]\Bigg),$$

where $\psi$ is the polygamma function,

$$\frac{\partial^2 \ell}{\partial \alpha^2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Bigg(I_{(y_{ij}=0)}\Bigg[\frac{(1-p_i)\left(\frac{\mu_i^{1+c}}{\alpha\left(\left(1+\alpha\mu_i^c\right)\right)^2}+\frac{2\mu_i}{\alpha^2\left(1+\alpha\mu_i^c\right)}-\frac{2\mu_i^{1-c}ln\left(1+\alpha\mu_i^c\right)}{\alpha^3}\right)}{\left(\left(1+\alpha\mu_i^c\right)\right)^{\frac{\mu_i^{1-c}}{\alpha}}\left(\frac{1-p_i}{\left(\left(1+\alpha\mu_i^c\right)\right)^{\frac{\mu_i^{1-c}}{\alpha}}}+p_i\right)}-$$

$$\frac{\left((1-p_i)\right)^2\left(\left(-\left(\frac{\mu_i}{\alpha\left(1+\alpha\mu_i^c\right)}\right)+\frac{\mu_i^{1-c}ln\left(1+\alpha\mu_i^c\right)}{\alpha^2}\right)\right)^2}{\left(\left(1+\alpha\mu_i^c\right)\right)^{\frac{2\mu_i^{1-c}}{\alpha}}\left(\left(\frac{1-p_i}{\left(\left(1+\alpha\mu_i^c\right)\right)^{\frac{\mu_i^{1-c}}{\alpha}}}+p_i\right)\right)^2}+$$

$$\frac{(1-p_i)\left(\left(-\left(\frac{\mu_i}{\alpha\left(1+\alpha\mu_i^c\right)}\right)+\frac{\mu_i^{1-c}ln\left(1+\alpha\mu_i^c\right)}{\alpha^2}\right)\right)^2}{\left(\left(1+\alpha\mu_i^c\right)\right)^{\frac{\mu_i^{1-c}}{\alpha}}\left(\frac{1-p_i}{\left(\left(1+\alpha\mu_i^c\right)\right)^{\frac{\mu_i^{1-c}}{\alpha}}}+p_i\right)}\Bigg]+I_{(y_{ij}>0)}\Bigg[$$

$$\frac{1}{\alpha^4\mu_i^{2c}\left(\left(1+\alpha\mu_i^c\right)\right)^2}\Big[2\alpha^2\mu_i^{1+2c}+3\alpha^3\mu_i^{1+3c}-\alpha^2\mu_i^{2c}y_{ij}-2\alpha^3\mu_i^{3c}y_{ij}-$$

$$2\alpha\mu_i^{1+c}ln\left(1+\alpha\mu_i^c\right)-4\alpha^2\mu_i^{1+2c}ln\left(1+\alpha\mu_i^c\right)-2\alpha^3\mu_i^{1+3c}ln\left(1+\alpha\mu_i^c\right)-$$

$$2\alpha\mu_i^{1+c}\left(\left(1+\alpha\mu_i^c\right)\right)^2\psi^{(0)}\left(\frac{\mu_i^{1-c}}{\alpha}\right)+2\alpha\mu_i^{1+c}\left(\left(1+\alpha\mu_i^c\right)\right)^2\psi^{(0)}\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)$$

$$-\mu_i^2\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right)-2\alpha\mu_i^{2+c}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right)-\alpha^2\mu_i^{2+2c}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}\right)+$$

$$\mu_i^2\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)+2\alpha\mu_i^{2+c}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)+\alpha^2\mu_i^{2+2c}\psi^{(1)}\left(\frac{\mu_i^{1-c}}{\alpha}+y_{ij}\right)\Big]\Bigg);$$

$$\frac{\partial^2 \ell}{\partial p_i \partial \alpha}=\frac{\partial^2 \ell}{\partial \alpha \partial p_i}=-\sum_{j=1}^{n_i}\Bigg(I_{(y_{ij}=0)}\Bigg[\frac{\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(\frac{\mu_i^{1-c}ln\left(1+\alpha\mu_i^c\right)}{\alpha^2}-\frac{\mu_i}{\alpha\left(1+\alpha\mu_i^c\right)}\right)}{(1-p_i)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}+p_i}+$$

$$\frac{(1-p_i)\left(1-\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}\right)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(\frac{\mu_i^{1-c}ln\left(1+\alpha\mu_i^c\right)}{\alpha^2}-\frac{\mu_i}{\alpha\left(1+\alpha\mu_i^c\right)}\right)}{\left((1-p_i)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}+p_i\right)^2}\Bigg]\Bigg);$$

$$\frac{\partial^2 \ell}{\partial p_i \partial \mu_i}=\frac{\partial^2 \ell}{\partial \mu_i \partial p_i}=-\sum_{j=1}^{n_i}\Bigg(I_{(y_{ij}=0)}\Bigg[\frac{\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(-\frac{c}{1+\alpha\mu_i^c}-\frac{(1-c)\mu_i^{-c}ln\left(1+\alpha\mu_i^c\right)}{\alpha}\right)}{(1-p_i)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}+p_i}+$$

$$\frac{(1-p_i)\left(1-\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}\right)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}\left(-\frac{c}{1+\alpha\mu_i^c}-\frac{(1-c)\mu_i^{-c}ln\left(1+\alpha\mu_i^c\right)}{\alpha}\right)}{\left((1-p_i)\left(1+\alpha\mu_i^c\right)^{-\frac{\mu_i^{1-c}}{\alpha}}+p_i\right)^2}\Bigg]\Bigg);$$

For the ZINB1, one obtains

$$\frac{\partial^2 \ell}{\partial \mu_i \partial \alpha} = \frac{\partial^2 \ell}{\partial \alpha \partial \mu_i} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \Big( I_{(y_{ij}=0)} \Big[ -(p_i - 1) \\ \frac{(1+\alpha) ln(1+\alpha) - \alpha)(\alpha p_i((1+\alpha)^{\frac{\mu_i}{\alpha}} - 1) - \mu_i p_i (1+\alpha)^{\frac{\mu_i}{\alpha}} ln(1+\alpha) + \alpha)}{\alpha^3 (1+\alpha)(p_i((1+\alpha)^{\frac{\mu_i}{\alpha}} - 1) + 1)^2} \Big] \\ + I_{(y_{ij}>0)} \Big[ \\ \frac{-\alpha \psi^{(0)} \left( \frac{\mu_i}{\alpha} + y_{ij} \right) + \mu_i (\psi^{(1)} \left( \frac{\mu_i}{\alpha} \right) - \psi^{(1)} \left( \frac{\mu_i}{\alpha} + y_{ij} \right)) + \alpha \psi^{(0)} \left( \frac{\mu_i}{\alpha} \right) + \alpha (\frac{1}{1+\alpha} + ln(1+\alpha) - 1))}{\alpha^3} \Big],$$

and for the ZINB2

$$\frac{\partial^2 \ell}{\partial \mu_i \partial \alpha} = \frac{\partial^2 \ell}{\partial \alpha \partial \mu_i} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \Big( I_{(y_{ij}=0)} \Big[ (1 - p_i) \\ \frac{\alpha^2 \mu_i (1-p_i) + p_i (1+\alpha\mu_i)^{\frac{1}{\alpha}} (\mu_i (\alpha^2 + \alpha(1 - ln(1+\alpha\mu_i))) - ln(1+\alpha\mu_i))}{\alpha^2 (1+\alpha\mu_i)^2 (p_i((1+\alpha\mu_i)^{\frac{1}{\alpha}} - 1) + 1)^2} \Big] \\ + I_{(y_{ij}>0)} \Big[ \frac{\mu_i - y_{ij}}{(1+\alpha\mu_i)^2} \Big].$$

.

Provided one has good starting values for $\hat{\boldsymbol{\beta}}^{(0)}$, $p^{(0)}$ and $\alpha^{(0)}$, the estimates for $\boldsymbol{\beta}$, $_i$ and $\alpha$ at the $(m+1)$th iteration are given by

$$\begin{pmatrix} \boldsymbol{\beta}^{(m+1)} \\ p^{(m+1)} \\ \alpha^{(m+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^{(m)} \\ p^{(m)} \\ \alpha^{(m)} \end{pmatrix} + \left[ \mathrm{I}^{(m)}(\boldsymbol{\beta}, p, \alpha) \right]^{-1} \mathrm{Sc}^{(m)}(\boldsymbol{\beta}, p, \alpha),$$

where the elements of I make use of the following expressions (from the ZINB pmf)

$$\mathrm{E}\left[ I_{(y_{ij}=0)} \right] = p_i + (1 - p_i)(1 + \alpha \mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}},$$

$$\mathrm{E}\left[ I_{(y_{ij}>0)} \right] = (1 - p_i) \left( 1 - (1 + \alpha \mu_i^c)^{-\frac{\mu_i^{1-c}}{\alpha}} \right).$$

## A.4 ZIP and ZINB MLE in the absence of covariates

Given observations $(y_1, ..., y_j)$ for $j = 1, ..., n$ such that $s = \sum_{j=1}^{n} y_j = n\bar{y}$, the resulting score equation for $p$ in the ZIP case becomes

$$\frac{\partial \ell}{\partial \mu} = -\frac{n_0(1-p)e^{-\mu}}{p + (1-p)e^{-\mu}} - (n - n_0) + \frac{s}{\mu}, \tag{A.4.1}$$

and

$$\frac{\partial \ell}{\partial p} = \frac{n_0(1-e^{-\mu})}{p + (1-p)e^{-\mu}} - \frac{n - n_0}{1-p}. \tag{A.4.2}$$

Setting each of these to zero yields

$$\frac{n_0(1-\hat{p})e^{-\hat{\mu}}}{\hat{p} + (1-\hat{p})e^{-\hat{\mu}}} + (n - n_0) = \frac{s}{\hat{\mu}}, \tag{A.4.3}$$

and

$$\frac{n_0(1-e^{-\hat{\mu}})}{p + (1-\hat{p})e^{-\hat{\mu}}} = \frac{n - n_0}{1-\hat{p}}. \tag{A.4.4}$$

Substituting (A.4.4) into (A.4.3) gives

$$\begin{aligned}
\frac{e^{-\hat{\mu}}(n - n_0)}{1 - e^{-\hat{\mu}}} + (n - n_0) &= \frac{s}{\hat{\mu}} \\
(n - n_0)\left(\frac{e^{-\hat{\mu}} + 1 - e^{-\hat{\mu}}}{1 - e^{-\hat{\mu}}}\right) &= \frac{s}{\hat{\mu}} \\
\frac{\hat{\mu}}{1 - e^{-\hat{\mu}}} &= \frac{s}{n - n_0}
\end{aligned}$$

$$\hat{\mu} = \frac{s\left(1 - e^{-\hat{\mu}}\right)}{n - n_0}, \tag{A.4.5}$$

thus independent of $p$. From (A.4.4), one has

$$\begin{aligned}
n_0(1 - e^{-\hat{\mu}}) - \hat{p}n_0(1 - e^{-\hat{\mu}}) &= \hat{p}(n - n_0) + (1 - \hat{p})(n - n_0)e^{-\hat{\mu}} \\
\hat{p}(n - n_0) - \hat{p}(n - n_0)e^{-\hat{\mu}} + \hat{p}n_0(1 - e^{-\hat{\mu}}) &= n_0(1 - e^{-\hat{\mu}}) - (n - n_0)e^{-\hat{\mu}},
\end{aligned}$$

giving

$$\hat{p} = \frac{n_0 - ne^{-\hat{\mu}}}{n(1 - e^{-\hat{\mu}})}. \tag{A.4.6}$$

For the ZINB model, the score equation for $p$ becomes

$$\frac{\partial \ell}{\partial p} = \frac{n_0 \left(1 - (1 + \alpha\mu^c)^{\frac{-\mu^{1-c}}{\alpha}}\right)}{(1-p)(1 + \alpha\mu^c)^{\frac{-\mu^{1-c}}{\alpha}} + p} - \frac{n - n_0}{1-p}. \tag{A.4.7}$$

Equating (A.3.8) to zero gives the MLE

$$\hat{p} = \frac{n(1 + \hat{\alpha}\hat{\mu}^c)^{\frac{-\hat{\mu}^{1-c}}{\hat{\alpha}}} - n_0}{n\left((1 + \hat{\alpha}\hat{\mu}^c)^{\frac{-\hat{\mu}^{1-c}}{\hat{\alpha}}} - 1\right)}. \tag{A.4.8}$$

The MLE of $\mu$ can be found by substituting (A.4.8) into (4.3.6) which provides the expression

$$\hat{\mu} = \frac{s\left(1 - (1 + \hat{\alpha}\hat{\mu}^c)^{\frac{-\hat{\mu}^{1-c}}{\hat{\alpha}}}\right)}{n - n_0}, \tag{A.4.9}$$

again which is independent of $p$. The remaining derivatives for the ZINB are as for covariate case in A.3 without the $i$ subscript.

The estimated covariance of $\hat{\mu}$ and $\hat{p}$, $\mathrm{Cov}(\hat{\mu}, \hat{p})$, can be found using the Fisher information matrix I, that is $\mathrm{Cov}(\hat{\mu}, \hat{p}) = \mathrm{I}^{-1}(\hat{\mu}, \hat{p})$, where

$$\mathrm{I}(\hat{\mu}, \hat{p}) = \begin{pmatrix} \mathrm{I}_{\mu\mu} & \mathrm{I}_{\mu p} \\ \\ \mathrm{I}_{p\mu} & \mathrm{I}_{pp} \end{pmatrix} \Big|_{\mu=\hat{\mu}, p=\hat{p}}.$$

The elements $\mathrm{I}_{\mu\mu}, \mathrm{I}_{\mu p} = \mathrm{I}_{p\mu}$ and $\mathrm{I}_{pp}$ are computed respectively via

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \mu^2}\right], -\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \mu p}\right] \text{and} - \mathrm{E}\left[\frac{\partial^2 \ell}{\partial p^2}\right].$$

For the ZIP, the second derivatives are

$$\frac{\partial^2 \ell}{\partial \mu^2} = \frac{n_0 p(1-p)e^{-\mu}}{(p + (1-p)e^{-\mu})^2} - \frac{s}{\mu^2};$$

$$\frac{\partial^2 \ell}{\partial \mu p} = \frac{\partial^2 \ell}{\partial p \mu} = \frac{n_0 e^{-\mu}}{(p + (1-p)e^{-\mu})^2};$$

$$\frac{\partial^2 \ell}{\partial p^2} = -\frac{n_0(1 - e^{-\mu})^2}{(p + (1-p)e^{-\mu})^2} - \frac{n - n_0}{(1-p)^2}.$$

Given that

$$\mathrm{E}\left[I_{(y_{ij}=0)}\right] = P(Y_{ij}=0) = p_i + (1-p_i)e^{-\mu_i},$$

$$\mathrm{E}\left[I_{(y_{ij}>0)}\right] = P(Y_{ij}>0) = (1-p_i)(1-e^{-\mu_i})$$

we have

$$\mathrm{I}_{\mu\mu} = -\mathrm{E}\left[\frac{\partial^2\ell}{\partial\mu^2}\right] = n\left[\frac{(1-p)}{\mu} - \frac{p(1-p)e^{-\mu}}{p+(1-p)e^{-\mu}}\right];$$

$$\mathrm{I}_{\mu p} = \mathrm{I}_{p\mu} = -\mathrm{E}\left[\frac{\partial^2\ell}{\partial\mu p}\right] = \frac{\partial^2\ell}{\partial p\mu} = -\frac{ne^{-\mu}}{p+(1-p)e^{-\mu}};$$

$$\mathrm{I}_{pp} = -\mathrm{E}\left[\frac{\partial^2\ell}{\partial p^2}\right] = \frac{n(1-e^{-\mu})}{(1-p)(p+(1-p)e^{-\mu})}.$$

$\mathrm{Var}(\hat{\mu})$ and $\mathrm{Var}(\hat{p})$ can be extracted from $\mathrm{I}^{-1}(\hat{\mu},\hat{p})$, using the inverse of the partitioned matrix as follows

$$\mathrm{Var}(\hat{\mu}) = (\mathrm{I}_{\mu\mu} - \mathrm{I}_{\mu p}\mathrm{I}_{pp}^{-1}\mathrm{I}_{p\mu})^{-1}\Big|_{\mu=\hat{\mu},p=\hat{p}} \tag{A.4.10}$$

$$\mathrm{Var}(\hat{p}) = (\mathrm{I}_{pp} - \mathrm{I}_{\mu p}\mathrm{I}_{\mu\mu}^{-1}\mathrm{I}_{p\mu})^{-1}\Big|_{\mu=\hat{\mu},p=\hat{p}}. \tag{A.4.11}$$

In our case, these quantities become

$$\begin{aligned}(\mathrm{I}_{\mu\mu} - \mathrm{I}_{\mu p}\mathrm{I}_{pp}^{-1}\mathrm{I}_{p\mu})\Big|_{\mu=\hat{\mu},p=\hat{p}} = n(1-\hat{p})\Big[&\frac{(\hat{p}+(1-\hat{p})e^{-\hat{\mu}})(1-e^{-\hat{\mu}}) - \hat{\mu}e^{-2\hat{\mu}}}{\hat{\mu}(\hat{p}+(1-\hat{p})e^{-\hat{\mu}})(1-e^{-\hat{\mu}})}\\ &- \frac{\hat{\mu}(1-e^{-\hat{\mu}})\hat{p}e^{-\hat{\mu}}}{\hat{\mu}(\hat{p}+(1-\hat{p})e^{-\hat{\mu}})(1-e^{-\hat{\mu}})}\Big]\end{aligned}$$

$$\begin{aligned}(\mathrm{I}_{pp} - \mathrm{I}_{\mu p}\mathrm{I}_{\mu\mu}^{-1}\mathrm{I}_{p\mu})\Big|_{\mu=\hat{\mu},p=\hat{p}} = n\Big[&\frac{(1-e^{-\hat{\mu}})}{(1-\hat{p})(\hat{p}+(1-\hat{p})e^{-\hat{\mu}})}\\ &- \frac{\hat{\mu}e^{-2\hat{\mu}}}{(1-\hat{p})(\hat{p}+(1-\hat{p})e^{-\hat{\mu}})(\hat{p}+(1-\hat{p})e^{-\hat{\mu}}-\hat{p}\hat{\mu}e^{-\hat{\mu}})}\Big].\end{aligned}$$

Since $\hat{p}+(1-\hat{p})e^{-\hat{\mu}} = \frac{n_0}{n}$ and $1-\hat{p} = \frac{\hat{\lambda}}{\hat{\mu}} = \frac{\hat{\bar{y}}}{\hat{\mu}}$, the above simplifies to

$$\mathrm{Var}(\hat{\mu}) = \frac{n_0\hat{\mu}^2(1-e^{-\hat{\mu}}))}{n\bar{y}\left[(1-e^{-\hat{\mu}})(n_0 - n(\hat{\mu}-\bar{y})e^{-\hat{\mu}}) - n\hat{\mu}e^{-2\hat{\mu}}\right]} \tag{A.4.12}$$

and

$$\mathrm{Var}(\hat{p}) = \frac{n_0\bar{y}(n_0 - n(\hat{\mu} - \bar{y})e^{-\hat{\mu}})}{n^2\hat{\mu}\left[(1-e^{-\mu})(n_0 - n(\hat{\mu} - \bar{y})e^{-\hat{\mu}}) - n\hat{\mu}e^{-2\hat{\mu}}\right]} \tag{A.4.13}$$

For the ZINBc, the following code can be used to estimate the 3x3 covariance matrix $\mathrm{I}^{-1}(\hat{\mu}, \hat{p}, \hat{\alpha})$:

```
zinbmatrix <- function(c,mu,alpha,p,ydata) {

  n <- length(ydata)
  expy0 <- n*(p+ (1-p)*(1+alpha*mu^(c))^(-mu^(1-c) /alpha))
  expy1 <- n*((1-p)*(1-(1+alpha*mu^(c))^(-mu^(1-c) /alpha)))
  expy <- n*mu*(1-p)

  Ipp <- expy1*(1 - p)^(-2) + expy0*(1 - (1 + alpha*mu^c)^(-(mu^(1 - c)
  /alpha)))^2 /((1 - p)*(1 + alpha*mu^c)^(-mu^(1 - c)/alpha) + p)^2
  Ipm <- Imp <- expy0*(((1 + alpha*mu^c)^(-mu^(1 - c) /alpha) *(-c/
  (1 + alpha*mu^c) - ((1 - c) *mu^(-c) *log(1 + alpha*mu^c))/alpha))/
  ((1 - p)*(1 + alpha*mu^c)^(-mu^(1 - c)/alpha) + p) + ((1 - p) *(1 -
  (1 + alpha*mu^c)^(-mu^(1 - c)/ alpha)) * (1 + alpha*mu^c)^(-mu^(1 - c)/
  alpha) *(-c/(1 + alpha*mu^c) - ((1 - c) *mu^(-c) *log(1 + alpha*mu^c))
  /alpha))/ ((1 - p)*(1 + alpha*mu^c)^(-mu^(1 - c)/ alpha) + p)^2)
  Ipa <- Iap <- expy0*(((1 + alpha*mu^c)^(-mu^(1 - c)/alpha) *((mu^(1 - c)
  *log(1 + alpha*mu^c))/alpha^2 - mu/(alpha *(1 + alpha*mu^c))))/
  ((1 - p)*(1 + alpha*mu^c)^(-mu^(1 - c)/alpha) + p) + ((1 - p)*
  (1 - (1 + alpha*mu^c)^(-mu^(1 - c)/alpha)) *(1 + alpha*mu^c)
  ^(-mu^(1 - c)/alpha)* ((mu^(1 - c) *log(1 + alpha*mu^c))/alpha^2 -
  mu/(alpha*(1 + alpha*mu^c))))/((1 - p) *(1 + alpha*mu^c)^
  (-mu^(1 - c)/alpha) + p)^2)

  y1data <- ydata[ydata>0]
  if(c==0){
  sum = 0
  for (j in 1:length(y1data)){
    sum = sum - (psigamma(y1data[j]+ mu/alpha, deriv = 1))/alpha^2
  }
  Immprt1 <- expy1*(psigamma(mu/alpha, deriv = 1))/alpha^2
  Immprt2 <- -expy0*(p*(p-1)*(1+alpha)^(mu/alpha)*(log(1+alpha))^2)/
  ((alpha*p)*((1+alpha)^(mu/alpha) -1) +alpha)^2
  Imm <- sum  + Immprt1 + Immprt2
  }
  else {
  sum = 0
  for (j in 1:length(y1data)){
    sum = sum + y1data[j]/(mu^2 *(alpha*mu + 1)^2) + (2*alpha* y1data[j])
    /(mu* (alpha*mu + 1)^2)
  }
  Immprt2 <- expy0*(((p-1)*p*(1+alpha)^(mu/alpha) *(log(1+alpha))^2))/
```

```
(alpha*p*((1+alpha)^(mu/alpha) -1) +alpha)^2
Immprt3 <- -expy1*alpha/(alpha*mu + 1)^2
Imm <- sum + Immprt2 + Immprt3
}

if (c==0){
  sum = 0
  for (j in 1:length(y1data)){
    sum = sum - mu*(2*alpha*psigamma(y1data[j]+mu/alpha, deriv = 0) +
    mu*(psigamma(y1data[j] + mu/alpha, deriv = 1)))/alpha^4
  }
  Iaaprt1a <- -((1-p)^2 *(alpha+1)^(-(2*mu)/alpha) *((mu*log(alpha+1))/
  alpha^2 - mu/(alpha*(alpha+1)))^2)/((1-p)*(alpha+1)^(-mu/alpha) + p)^2
  Iaaprt1b <- ((1-p)*(alpha+1)^(-mu/alpha) *((mu*log(alpha+1))/alpha^2 -
  mu/(alpha*(alpha+1)))^2)/((1-p)*(alpha+1)^(-mu/alpha) + p)
  Iaaprt1c <- ((1-p)*(alpha+1)^(-mu/alpha) *(-(2*mu*log(alpha+1))/alpha^3
  + (2*mu)/(alpha^2 *(alpha+1)) + mu/(alpha*(alpha+1)^2)))/
  ((1-p)*(alpha+1)^(-mu/alpha) +p)
  Iaaprt1 <- -expy0*sum(Iaaprt1a+Iaaprt1b+Iaaprt1c)
  Iaaprt2 <- (1+2*alpha)*expy/(alpha^2 *(1+alpha)^2)
  Iaaprt3 <- expy1*(mu*(2*(1+alpha)^2 *log(1+alpha) -alpha*(3*alpha +2)))/
  ((alpha^3)*(1+alpha)^2)
  Iaaprt4 <- expy1*(mu*(mu*psigamma(mu/alpha, deriv = 1) +
  2*alpha*psigamma(mu/alpha, deriv = 0)))/alpha^4
  Iaa <- sum + Iaaprt1 + Iaaprt2 + Iaaprt3 + Iaaprt4
}
else {
sum = 0
  for (j in 1:length(y1data)){
    sum - (2*alpha*psigamma(1/alpha + y1data[j], deriv=0) +
    psigamma(1/alpha + y1data[j], deriv=1))/alpha^4 - y1data[j]/
    (alpha^4 *mu^2 *(1+1/(alpha*mu))^2) + 2*y1data[j]/(alpha^3 *mu
    *(1+1/(alpha*mu)))
  }
  Iaaprt1 <- -expy0*(-((1 - p)^2 *(alpha*mu + 1)^(-2/alpha) *(log(
  alpha*mu + 1)/alpha^2 - mu/(alpha *(alpha*mu + 1)))^2)/((1 - p)
  *(alpha*mu + 1)^(-1/alpha) + p)^2 + ((1 - p) *(alpha*mu + 1)^(-1/alpha)
  *(log(alpha*mu + 1)/alpha^2 - mu/(alpha *(alpha*mu + 1)))^2)/((1 - p)*
  (alpha*mu + 1)^(-1/alpha) + p) + ((1 - p) *(alpha*mu + 1)^(-1/alpha) *
  (-(2* log(alpha*mu + 1))/alpha^3 + (2*mu)/(alpha^2 *(alpha*mu + 1)) +
  mu^2/(alpha *(alpha*mu + 1)^2)))/((1 - p)* (alpha*mu + 1)^(-1/alpha) + p))
  Iaaprt2 <- -expy1*(-2*alpha*psigamma(1/alpha, deriv=0) -
  psigamma(1/alpha, deriv=1))/alpha^4
  Iaaprt3 <- -expy1*(-2*log(1+ alpha*mu)/alpha^3 + 2*mu/(alpha^2 *
  (1+alpha*mu)) + mu^2 /(alpha*(1+alpha*mu)^2))
  Iaa <- sum + Iaaprt1 + Iaaprt2 + Iaaprt3
}
```

```
  if (c==0){
    sum = 0
    for (j in 1:length(y1data)){
      sum = sum + (mu*psigamma(y1data[j]+ mu/alpha, deriv = 1) +
      alpha*psigamma(y1data[j]+ mu/alpha, deriv = 0))/alpha^3
    }
    Iamprt1 <- -expy0*(1-p)*(alpha*(log(1+alpha)-1)+log(1+alpha))*
    (p*(1+alpha)^(mu/alpha) *(alpha- mu*log(1+alpha)) +alpha*(1-p))/
    (alpha^3 *(1+alpha)*(p*((1+alpha)^(mu/alpha) -1)+1)^2)
    Iamprt2 <- -expy1*(log(1+alpha)-(alpha/(1+alpha)))/alpha^2
    Iamprt3 <- -expy1*(mu*psigamma(mu/alpha, deriv = 1) +
    alpha*psigamma(mu/alpha, deriv = 0))/alpha^3
    Ima <- Iam <- sum + Iamprt1 + Iamprt2 + Iamprt3
  }
  else{
    sum = 0
    for (j in 1:length(y1data)){
      sum = sum +  y1data[j]/(1 + alpha*mu)^2
    }
    Iamprt1 <- -expy1*(mu/(1 + alpha*mu)^2)
    Iamprt2 <- -expy0*(((1-p) *(alpha^2 * mu*(1-p) + p*
    (alpha*mu + 1)^(1/alpha) *(mu *(alpha^2 + alpha*
    (1 - log(alpha*mu + 1))) - log(alpha *mu + 1))))/
    (alpha^2 *(alpha*mu + 1)^2 *(p *((alpha*mu + 1)^(1/alpha) - 1) + 1)^2))
    Ima <- Iam <- sum + Iamprt1 + Iamprt2
  }
  Hessian <- t(matrix(c(Imm,Imp,Ima,Ipm,Ipp,Ipa,Iam,Iap,Iaa),nrow = 3,
  ncol = 3))
  return(solve(Hessian))
}
zinbmatrix(c=0/1, \hat{\mu}, \hat{\alpha}, \hat{p}, data sample)[1,1]
zinbmatrix(c=0/1, \hat{\mu}, \hat{\alpha}, \hat{p}, data sample)[2,2]
```

## A.5 The delta method

For a vector of parameter estimates $\boldsymbol{\theta}$, let us assume a real-valued function $h(\boldsymbol{\theta})$. The multivariate Taylor Expansion tells us that $h(\boldsymbol{\theta}) : \mathbb{R}^w \to \mathbb{R}$ for $\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \\ \theta_w \end{pmatrix}$. Given the MLE, $\hat{\boldsymbol{\theta}}$, with $\text{Var}(\hat{\boldsymbol{\theta}}) = \Sigma$, applying Taylor's theorem to $h(\hat{\boldsymbol{\theta}})$ provides

$$h(\hat{\boldsymbol{\theta}}) = h(\boldsymbol{\theta}) + \nabla h(\boldsymbol{\theta})'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

$$\begin{aligned} \text{Var}(h(\hat{\boldsymbol{\theta}})) &= \nabla h(\boldsymbol{\theta})' \text{Var}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \nabla h(\boldsymbol{\theta}) \\ &= \nabla h(\boldsymbol{\theta})' \text{Var}(\hat{\boldsymbol{\theta}}) \nabla h(\boldsymbol{\theta}) \\ &= \nabla h(\boldsymbol{\theta})' \Sigma \nabla h(\boldsymbol{\theta}). \end{aligned}$$

Using the estimated gradients as approximations, one has

$$\begin{aligned} \text{Var}(h(\hat{\boldsymbol{\theta}})) &= \nabla h(\hat{\boldsymbol{\theta}})' \Sigma \nabla h(\hat{\boldsymbol{\theta}}) \\ &= \left( \frac{\partial h}{\partial \theta_1}, ..., \frac{\partial h}{\partial \theta_w} \right) |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \begin{pmatrix} \Sigma_{11} & * \\ \\ * & \Sigma_{ww} \end{pmatrix} \begin{pmatrix} \frac{\partial h}{\partial \theta_1} \\ \\ \frac{\partial h}{\partial \theta_w} \end{pmatrix} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \end{aligned}$$

Assuming the covariance components have very small magnitude [6], i.e. $\Sigma_{ij} \approx 0$ for $i \neq j$, the above expression reduces to

$$\begin{aligned} \text{Var}(h(\hat{\boldsymbol{\theta}})) &= \nabla h(\hat{\boldsymbol{\theta}})' \Sigma \nabla h(\hat{\boldsymbol{\theta}}) \\ &= \left( \frac{\partial h}{\partial \theta_1}, ..., \frac{\partial h}{\partial \theta_w} \right) |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \begin{pmatrix} \Sigma_{11} \frac{\partial h}{\partial \theta_1} \\ \\ \Sigma_{ww} \frac{\partial h}{\partial \theta_w} \end{pmatrix} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= \sum_{j=1}^{w} \left( \frac{\partial h}{\partial \theta_j} \right)^2 |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \Sigma_{jj} \end{aligned}$$

which is the multibiodose simplifcation. In our context, where $\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\mu} \end{pmatrix}$ and $h(\hat{\boldsymbol{\theta}})$ is the dose estimator $\hat{D}$, we have

$$\begin{aligned}\mathrm{Var}(\hat{D}) &= \sum_{j=1}^{3} \left(\frac{\partial h}{\partial \theta_j}\right)^2 |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \Sigma_{jj} \\ &= \left(\frac{\partial \hat{D}}{\partial \hat{\beta}_0}\right)^2 \Sigma_{\hat{\beta}_0\hat{\beta}_0} + \left(\frac{\partial \hat{D}}{\partial \hat{\beta}_1}\right)^2 \Sigma_{\hat{\beta}_1\hat{\beta}_1} + \left(\frac{\partial \hat{D}}{\partial \hat{\mu}}\right)^2 \Sigma_{\hat{\mu}\hat{\mu}}.\end{aligned}$$

It is almost certain that laboratories will not state the covariance terms with the reported calibration curve, therefore restricting the delta method to the non-covariance form.

## A.6 BfS data generation and cleaning

Blood was collected from healthy donors via an 18- or 20-gauge indwelling cannula (Vasofix Safety IV; B. Braun Melsungen AG, Melsungen, Germany) into 7.5ml lithium heparin monovettes (S-Monovette; Sarstedt AG & Co, Nümbrecht, Germany), mixed and portioned into 15ml centrifuge tubes (Falcon; Fisher Scientific GmbH) prior to irradiation on an X-ray high-protection device RS225 (195kV, 10mA, 0.5mm Cu filter, sample distance from X-ray tube 500 FSD, dose rate of 0.59Gy per minute, room temperature). All tubes were placed in the middle of the center in a horizontal position (X-Strahl Limited, UK). After irradiation, samples were incubated at 37°C for 60 min, kept at 5°C until isolation of peripheral blood leucocytes by density gradient centrifugation (10 min, 1000g, 5°C) using 12ml separation tubes (Leucosep Tube; Greiner Bio-One GmbH, Frickenhausen, Germany) and separation medium (Histopaque-1077; Sigma Aldrich Chemie GmbH, Taufkirchen, Germany). After centrifugation, leucocytes were transferred into 5ml cell culture medium (RPMI 1640; Pan-Biotech GmbH, Aidenbach, Germany). Cell suspension was centrifuged again (10 min, 250g, 5°C), and cells pellet was fixed in 2% paraformaldehyde (PFA; Sigma Aldrich) / phosphate buffered saline (Dulbecco's PBS; Biochrom GmbH, Berlin, Germany) solution for 15 min at 5°C before centrifugation (10 min, 250g, 5°C).

Lymphocytes were concentrated to one million cells per ml in PBS and stored at 5 °C. 100µl of cell suspension was spotted onto glass slides by cytospin centrifugation for 5 min at 54g. Slides were washed three times in fresh PBS containing 0.15% TritonX-100 (Sigma Aldrich) each time for 5 min, followed by three washing steps in blocking solution (1g bovine serum albumin (BSA; Sigma Aldrich) mixed with 0.15g glycine (Sigma Aldrich) in 100ml PBS each for 10 min. 75µl blocking solution with anti-phosphohistone H2A.X (Ser139) rabbit mAb (Cell Signaling Technology Europe B.V., Frankfurt a.M. Germany) in the dilution 1:200 was transferred on each slide and incubated at 4°°C for at least 16 hours. Slides were washed (5 min in PBS, for 10 min in PBS/Triton and for 5 min in PBS). Before incubating with the secondary antibody, an anti-rabbit IgG (H+L), F(ab')2 fragment conjugated to Alexa Fluor 555 fluorescent dye (Cell Signaling Technology Europe), in the dilution 1:1000 in blocking solution in a humid chamber for 45 min at room temperature slides were treated with blocking solution (7 min.). After antibody binding, slides were washed twice in PBS/Triton (5 min each), PBS (10 min and 7 min). Cell nuclei were counterstained with Hoechst 33342 (Bisbenzimide H 33342 trihydrochloride; Sigma Aldrich) for 2 min and slides were washed twice in PBS (2 min). Finally, slides were covered by 16µl antifade mounting medium (Vectashield; Vector Laboratories Inc., Burlingame, USA).

Search and image acquisition of cell nuclei on the slides was performed by automatic fluorescence microscopy using a scanning and imaging platform (Metafer 4, version V3.13.1; Meta-Systems Hard & Software GmbH, Altlussheim, Germany) equipped

with an objective (ZeissPlan-Neofluar 40× 0.75; Carl Zeiss Microscopy GmbH, Jena, Germany) yielding a 400-fold magnification. For foci analysis a Spectrum Orange bandpass filter (excitation: center wavelength/bandwidth = 546/10 nm, emission: 580/30 nm; Chroma 31003; Chroma Technology, Olching, Germany) and for counterstaining a DAPI bandpass filter (excitation: 350/50 nm, emission: 460/50 nm; Chroma 31000; Chroma Technology) was used. A foci specific Classifier 2.0.1 was created and used in all experiments.

The data set discussed in this paper is part of an even larger data set, consisting originally of 672 slides with a total of 1251882 foci counts, collected at the BfS in the six month period from July 2018 to January 2019. To arrive at the data presented here, all slides corresponding to any level of dose less than 0.1Gy were removed. In addition, the following cleaning steps had been carried out (post-scoring): (i) removed all slides with less than 800 foci counts, as a lower count indicates problems with the processing of the slide; (ii) removed slides which contained obvious data entry or measurement errors which could not be corrected; (iii) removed slides which were based on samples from a different experimental setup.

## A.7 Violation of QP independence

### A.7.1 Simulation

In Section 6.3, we verified through simulation the dependency effect (for a fixed covariate value dose) through three different heterogeneity cases. The R code to reproduce the results in Table 6.5 is presented below.

```
intercepts <- c(1,2)  # these are the two possible Poisson means
# lambda_1 and lambda_2
q <- c(0.5, 0.5) # probability q = 0.5
jmax = 1000
j<-1
yM <- matrix(0, 1000, jmax)
while (j <=jmax){
  # Run one of the following three commands:
  # (A) all Poisson means are independently chosen
  # r.intercepts <- sample(intercepts, 1000, replace=TRUE, prob=q)
  # (B) all Poisson means are the same for a fixed row
  # r.intercepts <- sample(intercepts, 1, replace=TRUE, prob=q)
  # (C)  within each row, strings of size 10 share the same
  # mean
  # r.intercepts <- rep(sample(intercepts, 10, replace=TRUE, prob=q),
  # each=100)
  xM <-rep(0,1000) # dose = 0
  yM[,j]<- rpois(1000,   r.intercepts) # generates Poisson counts
  j<-j+1
  if ((j %%10) ==0 ){print(j)}
}

# (A)
var(as.vector(yM))/mean(as.vector(yM)) # raw dispersion
# [1] 1.16772
var(colSums(yM))/mean(colSums(yM))   # aggregated dispersion
# [1] 1.070203

# (B)
var(as.vector(yM))/mean(as.vector(yM))
# [1]  1.168
var(colSums(yM))/mean(colSums(yM))
# [1] 168.9676

# (C)
var(as.vector(yM))/mean(as.vector(yM))
# [1] 1.166621
var(colSums(yM))/mean(colSums(yM))
# [1] 16.85397
```

### A.7.2 Theoretical derivation

We now present the theory behind the dispersion estimates for the two-component mixture model (6.3.1). We begin with deriving (6.3.5) and (6.3.6). Recall that $y_{ij}$ denotes the $j$-th count (cell) for slide $i$ with $j = 1, .., n$, and that $Z_{ij} \sim B(1, q)$, where yet no assumptions on the dependency structure of the $Z_{ij}$ are being made. Then,

$$\begin{aligned} E(y_{ij}) &= E(E(y_{ij}|Z_{ij})) \\ &= E(Z_{ij}\lambda_1 + (1 - Z_{ij})\lambda_2) \\ &= q\lambda_1 + (1 - q)\lambda_2 \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(y_{ij}) &= E(\mathrm{Var}(y_{ij}|Z_{ij})) + \mathrm{Var}(E(y_{ij}|Z_{ij})) \\ &= E({Z_{ij}}^2\lambda_1 + (1 - Z_{ij})^2\lambda_2) + \mathrm{Var}(Z_{ij}\lambda_1 + (1 - Z_{ij})\lambda_2) \\ &= q\lambda_1 + (1 - q)\lambda_2 + q(1 - q)(\lambda_1 - \lambda_2)^2. \end{aligned}$$

By dividing these two expressions, the dispersion index for the individual counts becomes (6.3.2). Now consider aggregated counts $s_i = \sum_{j=1}^{n} y_{ij}$. Then

$$E(s_i) = \sum_{i=1}^{n} E(y_{ij}) = n(q\lambda_1 + (1 - q)\lambda_2)$$

and

$$\begin{aligned} \mathrm{Var}(s_i) &= \mathrm{Var}\left(\sum_{i=1}^{n} y_{ij}\right) \\ &= \sum_{i=1}^{n} \mathrm{Var}(y_{ij}) + \sum_{j\neq l=1}^{n} \mathrm{Cov}(y_{ij}, y_{il}) \\ &= n(q\lambda_1 + (1 - q)\lambda_2 + q(1 - q)(\lambda_1 - \lambda_2)^2) + \sum_{j\neq l=1}^{n} \mathrm{Cov}(y_{ij}, y_{il}) \end{aligned}$$

which after division gives (6.3.7).

Consider now the special case $Z_{ij} \equiv Z_i$ (6.3.3). Then from (6.3.7) and (6.3.4),

$$\begin{aligned}
\phi_{agg} &= 1 + \frac{q(1-q)(\lambda_1-\lambda_2)^2}{q\lambda_1+(1-q)\lambda_2} + \frac{\sum_{j\neq l}^{n} q(1-q)(\lambda_1-\lambda_2)^2}{n(q\lambda_1+(1-q)\lambda_2)} \\
&= 1 + \frac{q(1-q)(\lambda_1-\lambda_2)^2}{q\lambda_1+(1-q)\lambda_2} + \frac{n(n-1)q(1-q)(\lambda_1-\lambda_2)^2}{n(q\lambda_1+(1-q)\lambda_2)} \\
&= 1 + \frac{nq(1-q)(\lambda_1-\lambda_2)^2}{q\lambda_1+(1-q)\lambda_2}
\end{aligned}$$

which proves (6.3.8).

Now assume the slide with $n$ cells consists of $b = \dfrac{n}{\tau}$ sub-groups (or strings) of size $\tau$, where all $y_{ij}$ in each batch are generated from the same distribution (either Pois($\lambda_1$) with probability $q$ or Pois($\lambda_2$) with probability $1-q$). (In terms of the experiment in Section 6.3.3, this setup corresponds to scenario (C) but covers scenario (B) in the case $\tau = n$, and scenario (A) in the case $\tau = 1$). This general model is hence formulated as

$$\begin{aligned}
y_{ij} &\sim Z_{ij}\text{Pois}(\lambda_1) + (1-Z_{ij})\text{Pois}(\lambda_2) \\
&= T_{ig}\text{Pois}(\lambda_1) + (1-T_{ig})\text{Pois}(\lambda_2)
\end{aligned}$$

where $j \in (\tau(g-1)+1, \tau g)$, $Z_{i,\tau(g-1)+1} = ... = Z_{i,sg} \equiv T_{i,g}$ and $T_{ig} \sim B(1,q)$ with $g = 1, ..., b$ independent; i.e., $g$ is the index of the subgroup.

The only required modification as compared to the previous derivation is to work out the covariances in the third term of (6.3.6). Observe here that the result (6.3.4) remains true but only for the observations within each string, that is

$$\text{Cov}(y_{ij}, y_{il}) = \begin{cases} q(1-q)(\lambda_1-\lambda_2)^2 & \text{if } j \text{ and } l \text{ from the same string;} \\ 0 & \text{otherwise.} \end{cases} \tag{A.7.1}$$

This implies

$$\begin{aligned}
\sum_{j\neq l} \text{Cov}(y_{ij}, y_{il}) &= \sum_{g=1}^{b} \sum_{j,l\in(\tau(g-1)+1,\tau g)}^{n} \text{Cov}(y_{ij}, y_{il}) \\
&= b\tau(\tau-1)q(1-q)(\lambda_1-\lambda_2)^2 \\
&= n(\tau-1)q(1-q)(\lambda_1-\lambda_2)^2
\end{aligned}$$

so that

$$\text{Var}(s_i) = n(q\lambda_1+(1-q)\lambda_2) + n\tau q(1-q)(\lambda_1-\lambda_2)^2.$$

Hence,

$$\begin{aligned}\phi_{agg} &= \frac{n(q\lambda_1 + (1-q)\lambda_2) + n\tau q(1-q)(\lambda_1 - \lambda_2)^2}{n(q\lambda_1 + (1-q)\lambda_2)} \\ &= 1 + \frac{\tau q(1-q)(\lambda_1 - \lambda_2)^2}{q\lambda_1 + (1-q)\lambda_2}\end{aligned}$$

which is just (6.3.9).

## A.8 CP/CNB applied to PHE-Foci1 samples

In relation to the estimates from the PHE-Foci2 dataset provided in Table 8.1, we notice from the reported $\hat{\alpha}$ in Table A.1 that the overdispersion in the PHE-Foci1 samples is mostly due to zero-inflation, most notably for exposures $\geq 60\%$. Consequently, this means that the CNB method was unable to improve on the dose and fraction estimates from the CP method in cases where $\hat{F}_{CNB} < F$.

| $F$ (%) | $D$ (Gy) | $\hat{\mu}_{CP}$ | $\hat{\mu}_{CNB}$ | $\hat{D}_{CP}$ | $\hat{D}_{CNB}$ | $\hat{F}_{CP}$ | $\hat{F}_{CNB}$ | $\hat{\alpha}$ |
|---|---|---|---|---|---|---|---|---|
| 30 | 0.75 | 2.712 | 2.664 | 1.145 | 1.117 | 0.509 | 0.518 | 0.144 |
| | | ±0.182 | ±0.164 | | | ±0.038 | ±0.042 | |
| | 1.5 | 2.962 | 2.912 | 1.292 | 1.263 | 0.511 | 0.520 | 0.168 |
| | | ±0.186 | ±0.170 | | | ±0.038 | ±0.040 | |
| | 3 | 4.381 | 4.312 | 2.126 | 2.086 | 0.476 | 0.484 | 0.439 |
| | | ±0.221 | ±0.240 | | | ±0.036 | ±0.037 | |
| 40 | 0.75 | 3.167 | 3.167 | 1.412 | 1.412 | 0.579 | 0.579 | 0.001 |
| | | ±0.178 | ±0.150 | | | ±0.037 | ±0.037 | |
| | 1.5 | 3.034 | 2.913 | 1.334 | 1.263 | 0.504 | 0.525 | 0.397 |
| | | ±0.189 | ±0.184 | | | ±0.037 | ±0.042 | |
| | 3 | 4.105 | 3.912 | 1.964 | 1.851 | 0.498 | 0.523 | 0.925 |
| | | ±0.210 | ±0.242 | | | ±0.036 | ±0.040 | |
| 60 | 0.75 | 3.059 | 3.059 | 1.349 | 1.349 | 0.624 | 0.624 | * |
| | | ±0.170 | ±0.097 | | | ±0.037 | ±0.037 | |
| | 1.5 | 3.200 | 3.200 | 1.432 | 1.432 | 0.589 | 0.589 | 0.001 |
| | | ±0.177 | ±0.168 | | | ±0.037 | ±0.037 | |
| | 3 | 4.790 | 4.790 | 2.367 | 2.367 | 0.565 | 0.565 | * |
| | | ±0.210 | ±0.222 | | | ±0.035 | ±0.035 | |
| 80 | 0.75 | 3.230 | 3.230 | 1.450 | 1.449 | 0.760 | 0.761 | * |
| | | ±0.157 | ±0.139 | | | ±0.033 | ±0.033 | |
| | 1.5 | 3.828 | 3.828 | 1.801 | 1.801 | 0.649 | 0.649 | 0.001 |
| | | ±0.179 | ±0.160 | | | ±0.035 | ±0.035 | |
| | 3 | 5.060 | 5.060 | 2.526 | 2.526 | 0.710 | 0.710 | 0.001 |
| | | ±0.192 | ±0.186 | | | ±0.032 | ±0.032 | |
| 100 | 0.75 | 3.093 | 3.093 | 1.369 | 1.369 | 0.895 | 0.895 | * |
| | | ±0.142 | ±0.034 | | | ±0.027 | ±0.027 | |
| | 1.5 | 4.301 | 4.301 | 2.080 | 2.080 | 0.816 | 0.816 | * |
| | | ±0.167 | ±0.118 | | | ±0.028 | ±0.028 | |
| | 3 | 5.537 | 5.537 | 2.807 | 2.807 | 0.934 | 0.934 | * |
| | | ±0.174 | ±0.099 | | | ±0.018 | ±0.018 | |

Table A.1 Dose and fraction estimates following procedures as outlined in Chapter 7 using the calibration curve $\lambda = 0.766 + 1.700D$. An asterisk * is used to indicate values $< 10^{-3}$.

# Bibliography


[1] Ainsbury, E. (2016). Dose and uncertainty estimation with the gamma-h2ax assay. unpublished factsheet. public health england.

[2] Ainsbury, E. and Lloyd, D. (2010). Dose estimation software for radiation biodosimetry. *Health Phys.*, 98:290–295.

[3] Ainsbury, E. A., Samaga, D., Della Monaca, S., Marrale, M., Bassinet, C., Burbidge, C. I., Correcher, V., Discher, M., Eakins, J., Fattibene, P., Güçlü, . I., Higueras, M., Lund, E., Maltar-Strmečki, N., McKeever, S., Rääf, C. L., Sholom, S., Veronese, I., Wieser, A., Woda, C., and Trompier, F. (2017). Uncertainty on radiation doses estimated by biological and retrospective physical methods. *Radiation Protection Dosimetry*, 178(4):382–404.

[4] Ainsbury et al (2013). Cytobayesj: Software tools for bayesian analysis of cytogenetic radiation dosimetry data. *Mutat Res*, 756:184–191.

[5] Ainsbury et al (2014). Review of bayesian statistical analysis methods for cytogenetic radiation biodosimetry, with a practical example. *Radiat Prot Dosimetry 2014*, 162:185–196.

[6] Ainsbury et al (2017). Uncertainty of fast biological radiation dose assessment for emergency response scenarios. *International Journal of Radiation Biology*, 93:1:127–135.

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

[8] Ali, Y Cucinotta, F., Ning-Ang, L., and Zhou, G. (2020). Cancer risk of low dose ionizing radiation. *Front Phys.*

[9] Amundson, S., Grace, M., McLeland, C., Epperly, M., Yeager, A., Zhan, Q., Greenberger, J., and Fornace, A. (2004). Human in vivo radiation-induced biomarkers: gene expression changes in radiotherapy patients. *Cancer Res*, 64(18):6368–6371.

[10] Anderson et al. (2013). Comparison of two methods for measuring h2ax nuclear fluorescence as a marker of dna damage in cultured human cells: Applications for microbeam radiation therapy. *Journal of Instrumentation.*

[11] Barnard, S., Bouffler, S., and Rothkamm, K. (2013). The shape of the radiation dose response for dna double-strand break induction and repair. *Genome Integrity*, 4: 1 pmid:23522792.

[12] Beels, L., Bacher, K., Smeets, P., Verstraete, K., Vral, A., and Thierens, H. (2011). Dose-length product of scanners correlates with dna damage in patients undergoing contrast ct. *Eur J Radiol.*

[13] Bender, M. and Gooch, P. (1962). Types and rates of x-ray-induced chromosome aberrations in human blood irradiated in vitro. *Proc Natl Acad Sci (USA)*, 48:522–532.

[14] Biostat (unknown). The gamma distribution and properties. *[Online]Available at: http://www.biostat.umn.edu/ cavanr/negBinDetailsV2.pdf.*

[15] Bohning, 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.

[16] Borras et al (2015). Comparison of methods to quantify histone h2ax phosphorylation and its usefulness for prediction of radiosensitivity. *International Journal of Radiation Biology*, pages 915–924.

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

[18] Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

[19] Byrum, S., Burdine, M., Orr, L., Mackintosh, S., Authier, S., Pouliot, M., Hauer-Jensen, M., and Tackett, A. (2017). Time- and radiation-dose dependent changes in the plasma proteome after total body irradiation of non-human primates: Implications for biomarker selection. *PLoS One*, 12(3):e0174771.

[20] Cahoy, D., Di Nardo, E., and Polito, F. (2021). Flexible models for overdispersed and underdispersed count data. *Statistical Papers*, 62:2969–2990.

[21] Cameron, C. and Trivedi, P. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, pages 29–54.

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

[23] Consul, P. (1989). Generalized poisson distributions: properties and applications. *Marcel Dekker, New York.*

[24] Dean, C. and Lawless, J. (1989). Tests for detecting overdispersion in poisson regression models. *Journal of the American Statistical Association*, 84(406):467–472.

[25] Deperas et al. (2007). Cabas: a freely available pc program for fitting calibration curves in chromosome aberration dosimetry. *Radiat. Prot. Dosim.*, 124:115–123.

[26] Dolphin, G. (1969). Biological dosimetry with particular reference to chromosome aberration analysis. *A review of methods. Handling of radiation accidents (Proc. Int. Symp. Vienna, 1969), IAEA, Vienna (1969)*, pages 215–224.

[27] Drugs Information, O. (2000). Drugs and diseases reference index. http://drugster.info/medic/term/chromosome-dicentric/.

[28] Einbeck, J., Ainsbury, E., Sales, R., Barnard, S., Kaestle, F., and Higueras, M. (2018). A statistical framework for radiation dose estimation with uncertainty quantification from the -h2ax assay. *PLoS ONE 13(11): e0207464.*

[29] Einbeck et al. (2015). On the use of random effect models for radiation biodosimetry. *Extended Abstracts Fall 2015. Trends in Mathematics. Basel*, 8:89–94.

[30] Errington, A., Einbeck, J., and Cumming, J. (2021a). Estimating exposure fraction from radiation biomarkers: A comparison of frequentist and bayesian approaches. *Advances in Uncertainty Quantification and Optimization Under Uncertainty with Aerospace Applications. UQOP 2020. Space Technology Proceedings. Springer, Cham.*, 8:393–405.

[31] Errington, A., Einbeck, J., and Cumming, J. (2021b). Radiation dose estimation via the contaminated poisson and negative binomial methods in partial-body exposures. *Mathematical and Statistical Methods for Metrology (MSMM) 2021 Booklet of Abstracts*.

[32] Errington, A., Einbeck, J., Cumming, J., Rössler, U., and Endesfelder, D. (2021c). The effect of data aggregation on dispersion estimates in count data models. *The International Journal of Biostatistics*, 18(1):183–202.

[33] Farrance, I. and Frenkel, R. (2014). Uncertainty in measurement: A review of monte carlo simulation using microsoft excel for the calculation of uncertainties through functional relationships, including uncertainties in empirically derived constants. *Clin Biochem Rev*, 35 (1).

[34] Fernández-Fontelo, A., Puig, P., Ainsbury, E., and Higueras, M. (2018). An exact goodness-of-fit test based on the occipancy problems to study zero-inflation and zero-deflation in biological dosimetry data. *Radiation Protection Dosimetry*, 179(4):317–326.

[35] Firsanov, D., Solovjeva, L., and Svetlova, M. (2011). H2ax phosphorylation at the sites of dna double-strand breaks in cultivated mammalian cells and tissues. *International Journal of Radiation Biology*, 2:283–297.

[36] Gao, Y. (2017). Gamma-h2ax-based dose estimation via standard methodology in dicentric assay. *Master of Science Dissertation. University of Durham.*

[37] Garty et al. (2010). The rabit: a rapid automated biodosimetry tool for radiological triage. *Health Phys.*, 98:209–217.

[38] Greene, W. (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*.

[39] Grudzenski, S., Schwab, S., Wiederseiner, M., Heckmann, M., Bautz, W., Lobrich, M., and Uder, M. (2009). Dna double-strand breaks and their repair in blood lymphocytes of patients undergoing angiographic procedures. *Invest Radiol*, page 44:440–446.

[40] Gunnar, B., Elisabeth, G., Ute, R., David, E., Alexandra, K., Ambros, B., and Matthias, E. (2020). Double-strand breaks in lymphocyte dna of humans exposed to [18f] fluorodeoxyglucose and the static magnetic field in pet/mri. *European Journal of Nuclear Medicine and Molecular Imaging*, (in press).

[41] Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, 2:e616.

[42] Hasnine, M et al (2018). Towards final scores prediction over clickstream using machine learning methods. Asia-Pacific Society for Computers in Education (APSCE).

[43] Hernandez et al (2019). https://aldomann.shinyapps.io/biodosetools-v3/.

[44] Higueras, M., Puig, P., Ainsbury, E., and Rothkamm, K. (2015a). Advanced statistical methods for cytogenetic radiation biodosimetry.

[45] Higueras, M., Puig, P., Ainsbury, E., and Rothkamm, K. (2015b). A new inverse regression model applied to radiation biodosimetry. *Proceedings of the Royal Society A*, pages 170–173.

[46] Higueras et al (2015). A new bayesian model applied to cytogenetic partial body irradiation estimation. *Radiat Prot Dosimetry 2015.*

[47] Hilali, A., Leonard, E., Decat, G., and Leonard, A. (1991). An appraisal of the value of the contaminated poisson method to estimate the dose inhomogeneity in simulated partial-body exposure. *Radiation Research*, 128:108–111.

[48] Hilbe, J. (2007). Negative binomial regression. *Cambridge University Press.*

[49] Hines, J. (1997). A comparison of score tests for overdispersion in generalized linear models. *Journal of Statistical Computation and Simulation*, pages 323–342.

[50] Hladik, D., Bucher, M., Endesfelder, D., and Oestreicher, U. (2022). The potential of omics in biological dosimetry. *Radiation*, 2:78–90.

[51] Holley, A., Miao, L., St. Clair, D., and St. Clair, W. (2014). Redox-modulated phenomena and radiation therapy: the central role of superoxide dismutases. *Antioxid Redox Signal*, 20(10):1567–1589.

[52] Horn, S., Barnard, S., and Rothkamm, K. (2011). Gamma-h2ax-based dose estimation for whole and partial body radiation exposure. *PLoS One*, page 6:e25113.

[53] IAEA (2011). Cytogenetic dosimetry: applications in preparedness for and response to radiation emergencies. *International Atomic Energy Agency: Vienna.*

[54] Joe, H. and Zhu, R. (1962). On the compound multinomial distribution, the multivariate $\beta$-distribution, and correlations among proportions. *Biometrika*, 49:65–82.

[55] Jones, J., Scott, A., Tudor, G., Xu, P., Jackson, I., Vujaskovic, Z., Booth, C., MacVittie, T., Ernst, R., and Kane, M. (2014). Identification and quantitation of biomarkers for radiation-induced injury via mass spectrometry. *Health Phys*, 106(1):106–119.

[56] Khoury, L., Zalko, D., and Audebert, M. (2020). Evaluation of the genotoxic potential of apoptosis inducers with the $\gamma$h2ax assay in human cells. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 852.

[57] Kinner, A., Wu, W., Staudt, C., and Iliakis, G. (2008). p53 binding protein 1 (53bp1) is an early participant in the cellular response to dna double-strand breaks. *Nucleic Acids Res*, pages 36(17):5678–94.

[58] Kopp, B., Khoury, L., and Audebert, M. (2019). Validation of the $\gamma$h2ax biomarker for genotoxicity assessment: a review. *Archives of Toxicology*, 93(8):2103–2114.

[59] Kuefner et al. (2010). Effect of ct scan protocols on x-ray-induced dna double-strand breaks in blood lymphocytes of patients undergoing coronary ct angiography. *Eur Radiol.*, 20:2917–2924.

[60] Kuo, L. and Yang, L. (2008). Gamma-H2AX – a novel biomarker for dna double-strand breaks. *in vivo*, 22:305–309.

[61] Laiakis, E., Pannkuk, E., Diaz-Rubio, M., Wang, Y., Mak, T., Simbulan-Rosenthal, C., Brenner, D., and Fornace, A. (2016). Implications of genotypic differences in the generation of a urinary metabolomics radiation signature. *Mutat Res*, 788:41–49.

[62] Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14.

[63] Lassmann et al, M. (2010). Eanm procedure guidelines for therapy of benign thyroid disease. *Eur J Nucl Med Mol Imaging 2010*, pages 37: 2218–2228.

[64] Lawless, J. (1987). Negative binomial and mixed poisson regressions. *The Canadian Journal of Statistics*, pages 209–225.

[65] Lin, T Tsai, M. (2013). Modeling health survey data with excessive zero and k responses. *Statistics in Medicine*, 32(9):1572–1583.

[66] Lloyd, D. and Edwards, A. (1983). Chromosome aberrations in human lymphocytes: effect of radiation quality, dose, and dose rate. *Radiation-Induced Chromosome Damage in Man. Alan R. Liss, New York, NY*, pages 23–49.

[67] Lloyd, D., Purrott, R., and Dolphin, G. (1973). Chromosome aberration dosimetry using human lymphocytes in simulated partial body irradiation. *Phys. Med. Biol*, 18:421–431.

[68] Lobrich, M., Rief, N., Kuhne, M., Heckmann, M., Fleckenstein, J., Rube, C., and Uder, M. (2005). In vivo formation and repair of dna double-strand breaks after computed tomography examinations. *Proc Natl Acad Sci USA.*, 102:8984–8989.

[69] Lu, T., Hsu, Y., Lai, L., Tsai, M., and Chuang, E. (2014). Identification of gene expression biomarkers for predicting radiation exposure. *Sci Rep*, 4(1):6293.

[70] Mandina, T., Roch-Lefèvre, S., Voisin, P., González, J., Lamadrid, A., Romero, I., Garcia, O., and Roy, L. (2011). Dose-response relationship of gamma-h2ax foci induction in human lymphocytes after x-rays exposure. *Radiation Measurements*, page 46:997–999.

[71] McCullagh, P. and Nelder, J. (1989). Generalized linear models. *Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series Chapman Hall (Second Edition).*

[72] Melkersson, M. and Olsson, C. (1999). Is visiting the dentist a good habit? analyzing count data with excess zeros and excess ones. *Umea Economic Studies*, 492:1–18.

[73] Merkle, W. (1983). Statistical methods in regression and calibration analysis of chromosome aberration data. *Radiat Environ Biophys*, 21:217–233.

[74] Moquet, J., Barnard, S., Staynova, A., Lindholm, C., Monteiro Gil, O., Martins, V., Rossler, U., Vral, A., Vandevoorde, C., and Wojewodzka, M. (2017). The second gamma-h2ax assay inter-comparison exercise carried out in the framework of the european biodosimetry network (reneb). *Int J Radiat Biol*, 93(1):58–64.

[75] Moran, P. (1971). Maximum likelihood estimation in non-standard conditions. *Proceedings of the Cambridge Philosophical Society*, pages 441–450.

[76] Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P., and Rothkamm, K. (2016). Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal*, 58(2):259–279.

[77] Oliveira et al (2016). Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal*, 58:259–279.

[78] Ostheim, P., Amundson, S., Badie, C., and et al. (2021). Gene expression for biodosimetry and effect prediction purposes: promises, pitfalls and future directions – key session conrad 2021. *International Journal of Radiation Biology*, 98(5):843–854.

[79] O'Brien, G., Cruz-Garcia, L., Majewski, M., Grepl, J., Abend, M., Port, M., Tichy, A., Sirak, I., Malkova, A., and Donovan, E. (2018). Fdxr is a biomarker of radiation exposure in vivo. *Sci Rep*, 8:684.

[80] Pernot, E., Hall, J., Baatout, S., Benotmane, M., Blanchardon, E., and et al. (2012). Ionizing radiation biomarkers for potential use in epidemiological studies. *Mutat Res*, 751(2):258–286.

[81] Perry, P. and Wolff, S. (1974). New giemsa method for differential staining of sister chromatids. *Nature (London)*, 251:156–158.

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

[83] Pujol, M., Barrios, L., Puig, P., Caballin, M., and Barquinero, J. (2016). A new model for biological dose-assessment in cases of heterogeneous exposures to ionizing radiation. *Radiation Research*, 185(2):151–162.

[84] Rao, C. and Chakravarti, I. (1956). Some small sample tests of significance for a poisson distribution. biometrics. 12:264–282.

[85] Redon, C., Nakamura, A., and Martin et al., O. (2011). Recent developments in the use of -h2ax as a quantitative dna double-strand break biomarker. *Aging (Albany NY)*, pages 3, 168–174.

[86] Redon, C. and Nakamura et al, A. (2010). The use of gamma-h2ax as a biodosimeter for total-body radiation exposure in non-human primates. *Plos One*, page 5(11):e15544.

[87] Ridout, M., Demetrio, C., and Hinde, J. (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57:219–223.

[88] Roch-Lefèvre, S., Mandina, T., Voisin, P., Gaëtan, G., Mesa, J., Valente, M., Bonnesoeur, P., García, O., Voisin, P., and Roy, L. (2010). Quantification of gamma-h2ax foci in human lymphocytes: a method for biological dosimetry after ionizing radiation exposure. *Radiat Res*, pages 174(2):185–94.

[89] Rogakou, E., Pilch, D., Orr, A., Ivanova, V., and Bonner, W. (1998). Dna double-stranded breaks induce histone h2ax phosphorylation on serine 139. *J Biol Chem*, March 6:273(10):5858–68.

[90] Romm et al (2013). Auto scoring of dicentric chromosomes as a tool in large scale radiation accidents. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 756:174–183.

[91] Rothkamm, K., Balroop, S., and Shekhdar, J et al. (2007). Leukocyte dna damage after multi-detector row ct: a quantitative biomarker of low-level radiation exposure. *Radiology*, page 242:244–51.

[92] Rothkamm, K. and Horn, S. (2009). Gamma-h2ax as protein biomarker for radiation exposure. *Ann Ist Super Sanita*, page 265–71.

[93] Rothkamm, K., Horn, S., Scherthan, H., Rossler, U., De Amicis, A., Barnard, S., Kulka, U., Lista, F., Meineke, V., and Braselmann, H. (2013). Laboratory intercomparison on the gamma-h2ax foci assay. *Radiat Res*, 180(2):149–155.

[94] Rothkamm et al. (2013). Manual versus automated $\gamma$-h2ax foci analysis across five european laboratories: Can this assay be used for rapid biodosimetry in a large scale radiation accident? *Mutat Res*, 756:170–173.

[95] Sak, A., Grehl, S., and Erichsen et al., P. (2007). Gamma-h2ax foci formation in peripheral blood lymphocytes of tumor patients after local radiotherapy to different sites of the body: dependence on the dose-distribution, irradiated site and time from start of treatment. *Int J Radiat Biol*, page 83:639–652.

[96] Sales, R. (2019). An analysis of uncertainty for dose estimation through the gamma-h2ax assay. *Masters thesis, Durham University.*

[97] Sasaki, M. and Miyata, H. (1968). Biological dosimetry in atom bomb survivors. *Nature*, 220:1189–1193.

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

[99] Sproull, M., Camphausen, K., and Koblentz, G. (2017a). Biodosimetry: a future tool for medical management of radiological emergencies. *Health Secur*, 15(6):599–610.

[100] Sproull, M., Kramp, T., Tandle, A., Shankavaram, U., and Camphausen, K. (2017b). Multivariate analysis of radiation responsive proteins to predict radiation exposure in total-body irradiation and partial-body irradiation models. *Radiat Res*, 187(2):251–258.

[101] Szluinska, M., Edwards, A., and Lloyd, D. (2005). Statistical methods for biological dosimetry. *Health Protection Agency, Centre for Radiation, Chemical and Environmental Hazards, Radiation Protection Division.*

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

[103] Venkateswarlu et al. (2015). Mean frequency and relative fluorescence intensity measurement of $\gamma$-h2ax foci dose response in pbl exposed to $\gamma$-irradiation: An inter- and intra-laboratory comparison and its relevance for radiation triage. *Cytometry Part A*, 87A:1138–1146.

[104] Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika 61*, pages 439–447.

[105] Wilson, A., Grabowski, P., Elloway, J., Ling, S., Stott, J., and Doherty, A. (2021). Transforming early pharmaceutical assessment of genotoxicity: applying statistical learning to a high throughput, multi end point in vitro micronucleus assay. *Sci Rep*, 11(2535).

[106] Zhang, C. and Tian, G Ng, K. (2016). Properties of the zero-and-one inflated poisson distribution and likelihood-based inference methods. *Statistics and Its Interface.*, 9:11–32.