An Analysis of Uncertainty for Dose Estimation through the -H2AX Assay

RACHEL MARY SALES

An Analysis of Uncertainty for Dose Estimation through the γ-H2AX Assay

Rachel Sales

Supervisor: Dr Jochen Einbeck

Department of Mathematical Sciences

Durham University

March 2019

A dissertation submitted in fulfilment of the requirements for admission to the degree of MSc by Research in Mathematical Sciences at Durham University

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.

Statement of Copyright

Acknowledgements

I would like to thank Jochen Einbeck, my supervisor, for his invaluable help, advice, and guidance. I am also thankful for Elizabeth Ainsbury, Stephen Barnard, Manuel Higueras, and Felix Kaestle, whose knowledge and collaboration have proved vital to my work. As well as this, I would like to thank my friends, family, and everyone else who has supported me throughout the writing of this thesis.

Abstract

Currently, the primary biomarker for radiation biodosimetry is the dicentric chromosome. However, the γ-H2AX histone is an alternative assay that is less labour-intensive. Blood samples can be taken more quickly than for the dicentric biomarker, and a larger number of samples can be handled within a given time frame. In this thesis, we discuss several statistical techniques for how to handle scored γ-H2AX foci data. We then apply these techniques to two datasets from Public Health England, using one to demonstrate techniques, and the second to check that the dose-response curve calculation and dose estimation techniques work. Throughout we choose to fit quasi-Poisson models instead of Poisson ones to account for overdispersion present within the foci count data. After fitting both linear and quadratic dose-response curves we create controls to validate the curves, using a reference sampling ratio to scale them if necessary. By calculating the uncertainty we show why linear fits are preferable to quadratic ones. We finally compare our linear dose-response curves from both datasets for multiple timepoints with pre-existing ones from the literature to see how they compare and what conclusions can be drawn about dose-response curves for this assay in general.

Contents

Chapter 1

Introduction

In this day and age, the use of nuclear power and radioactive materials means that there is a risk to the public of accidental irradiation, should something go wrong. In a radiation incident, scientists need to be able to rapidly and reliably determine each individual victim's level of exposure and thus the dose that they have contracted. However, we have an issue, namely that we cannot guarantee that an exposed individual will be wearing a radiation dosimeter (members of the public typically do not) [1]. According to their review, 37% (27) of the cases investigated by Public Health England (PHE) between 2006 and 2015 were listed as "Suspected overexposure of people not wearing a dosimeter" [2]. The majority (59%, or 43 cases) of the 73 investigations during that decade were due to possible non-uniform exposure, specifically "that in which the relationship between dose to the physical dosimeter and to the body is uncertain". This sort of situation is one where an unexpectedly high dosimeter reading is recorded, and samples are taken as a precaution.

Biological dosimetry is useful because it can be used to initially split victims into two groups: those deemed to be "critically exposed", who should be prioritised, and the "worried well", people who have (comparatively) been minimally exposed thus are unlikely to need urgent treatment [3]. As well as this, it can provide us with useful information regarding the probable future health consequences, both stochastic and deterministic, for victims of radiation incidents [4].

1.1 Radiation Dose Modelling

The most commonly used type of biomarker is chromosomal aberrations, wrapped-up DNA that has been damaged or otherwise altered by ionising radiation.

1.2 The Dicentric Biomarker

Currently, the main aberration used in biological dosimetry is the dicentric chromosome. This is a chromosome with two centromeres ("crossings") instead of the usual one. 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 [5].

The dicentric chromosome is a cytogenetic biomarker, a type of biomarker that counts chromosomal aberrations in blood lymphocytes. Cytogenetic biomarkers have been considered to be the "gold standard" for radiation biodosimetry for three decades, mostly due to a comparative lack of inter-individual variation [6].

Due to the amount of research that's been done into it and the widespread utilisation of this method, 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 [7]. 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 [7]. This figure is as high as it is due to well developed international mutual assistance networks such as the EU funded project RENEB (Realising the European NEtwork of Biodosimetry) [8]. This presents us with a clear issue: should there be 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 that can be assessed sooner and more quickly.

1.3 The Public Health England Factsheet

In this project we will analyse an unpublished internal factsheet (2.5 pages) from Public Health England [9] entitled "Dose and uncertainty estimation with the gamma-H2AX assay". This factsheet briefly and succinctly discusses suggested procedures to estimate doses from blood or lymphocyte samples using the γ-H2AX assay biomarker. We will discuss and explore the techniques detailed on it with reference to the field of biodosimetry as a whole then apply them to two PHE datasets. The first dataset, henceforth referred to as PHE (2017), is our primary dataset, as it contains a greater number of design doses

for each time point (as well as a larger number of usable measurements) and was fully available when this thesis was started. The second dataset is known as PHE (2018) and is used to check that the methods described work for scenarios other than the first dataset.

Some of the material in this thesis fed into a paper that has been accepted for publication. (The author of this thesis is listed as a co-author of the paper.) This paper, entitled "A statistical framework for radiation dose estimation with uncertainty quantification from the γ-H2AX assay", has been published by the journal PLoS One, and any figures used from it will be suitably cited [10]. The aim of the paper is to establish a combined statistical methodology for calibration curve estimation, dose estimation, and uncertainty quantification for the γ-H2AX assay, as currently practical use of the assay is limited by a lack of agreement on which strategies to use. This contrasts with the dicentric assay, which has standardised procedures, as given in the IAEA manual (2011) [5].

1.4 Introducing the First Dataset

Our initial dataset contains 339 foci/cell measurements (also known as "yields"). 32 individual PHE staff volunteers provided blood samples, which were then irradiated ex-vivo with 250 kVp X-rays. For this dataset 7 design dose points were used: 0Gy, 0.05Gy, 0.1Gy, 0.25Gy, 0.5Gy, 1Gy, and 4Gy. Yields were recorded at only two time points after exposure, 1h and 24h, for groups of $n = 500$ cells.

Figure 1.1 is an initial scatterplot of the raw data, colour coded by timepoint. Here we have that the spread of data increases with dose, and that the 24h data has a smaller spread than the 1h data at every dose marker where comparison is possible. Excluding one potential (large) outlier, the 4Gy 24h data has a smaller spread than the 1Gy 1h data. There are clear positive trends within the data, although it is clear that the foci/cell count is higher overall and increases at a larger rate for the 1h data.

The number of measurements used per individual varies from 1 (person H62) to 32 (person H9), as shown in Figure 1.2. All of the models we fitted utilised data from multiple samples, choosing not to consider the sampling individual. This implies an assumption of inter-individual variation between foci counts that is not larger than intra-individual variation, something that is taken to be standard with regards to this area of biodosimetry and will be discussed further in Chapter 2. That being said, should an individual's sample results later be found to be compromised for some reason the label can be used to suitably remove those readings only from the dataset. Labelling the individuals (anonymously)

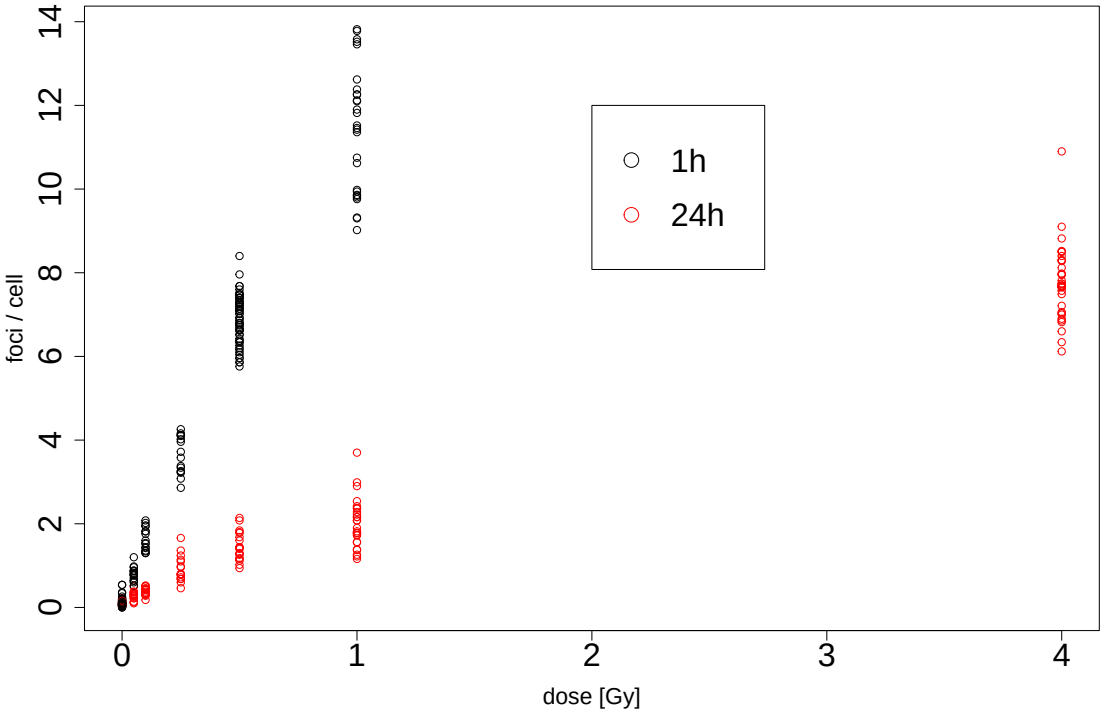

Figure 1.1: A scatterplot showing the raw data from the PHE (2017) dataset, colour coded by timepoint. The 1h data is shown in black, whilst the 24h data is shown in red.

also allows us to double check levels of inter-individual variation if we choose to do so, although we must be mindful of the wildly varying number of samples per person.

Of the 339 measurements in this dataset, 78 of them (just over 23% of the total) were control measurements, those taken at a dose point of 0Gy. Aside from those, the most common dose for samples to be exposed to was 0.5Gy (74 measurements, or approximately 21.8%). The three lowest exposed doses (0.05Gy, 0.1Gy, and 0.25Gy) tied with the least numbers of measurements at 32 each. For the highest three doses (0.5Gy, 1Gy, 4Gy) used in this dataset the number of readings decreases with dose by 19 each time (74, 55, 36). It must be noted, however, that the 4Gy data is all from a single timepoint (24h), whilst every other dose records at least one measurement from each of the two timepoints.

Figure 1.3 shows the number of readings per design dose for each timepoint, and it should be noted that the only dose where we have more 24h measurements than 1h ones is 4Gy, where we have no data for the 1h timepoint. We have equal numbers of readings per timepoint (16 per timepoint per dose) for three out of the seven stated doses (0.05Gy, 0.1Gy, 0.25Gy). It is noticeable that these are the three lowest doses that cells were irradiated with, and that (excluding the 0 recorded for the 1Gy and 4h combination) this

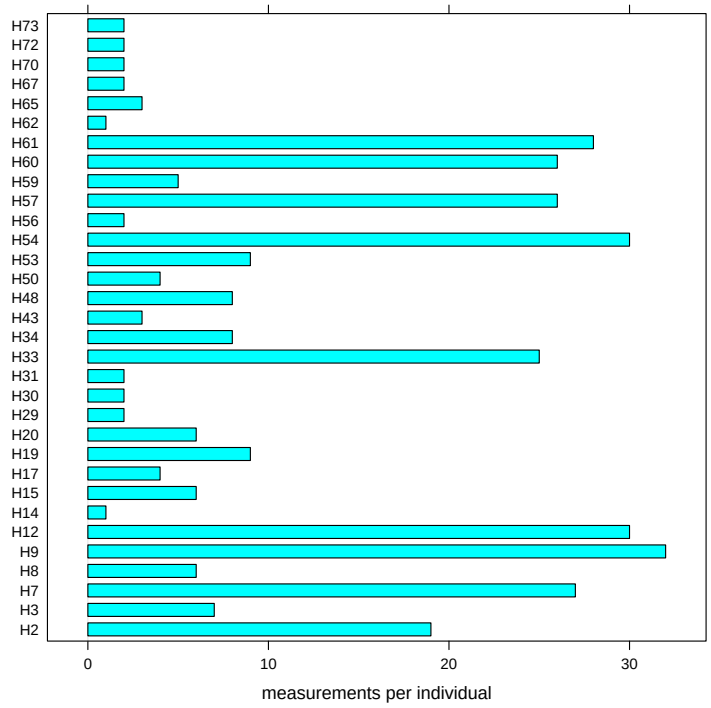

Figure 1.2: A bar chart showing the number of measurements (foci counts per 500 cells) taken per individual, sorted in order of anonymous user code. From [10].

number of measurements recorded for each dose and time pair is the joint lowest. This could potentially be due to difficulties working with very small doses. For the 1Gy data we almost have an equal number of readings per timepoint, with 28 for 1h and 27 for 24h. This suggests that there may have initially been an equal number of readings taken per timepoint, although if so clearly at least one was unusable. For the other two doses where we have readings for both timepoints (0Gy and 0.5Gy) we have over twice as many readings for the 1h data than for the 24h data. Of the 78 control measurements, only 22 of them were taken 24h after exposure. Nearly three-quarters of the 74 readings taken at 0.5Gy were taken 1h after exposure (55 were taken after 1h, the other 19 after 24h). The dose with the largest number of 24h readings is 4Gy, which has 36 of them.

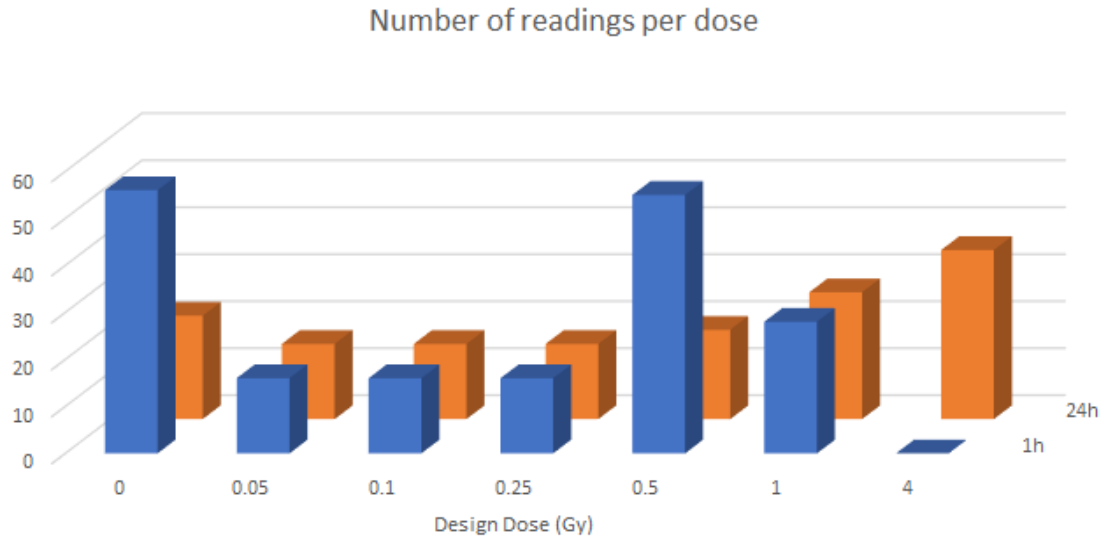

Figure 1.3: A bar chart showing the number of measurements recorded for each of the seven doses in the dataset, sorted by timepoint.

Chapter 2

The γ-H2AX Histone Biomarker

In this thesis we will focus on the γ-H2AX biomarker.

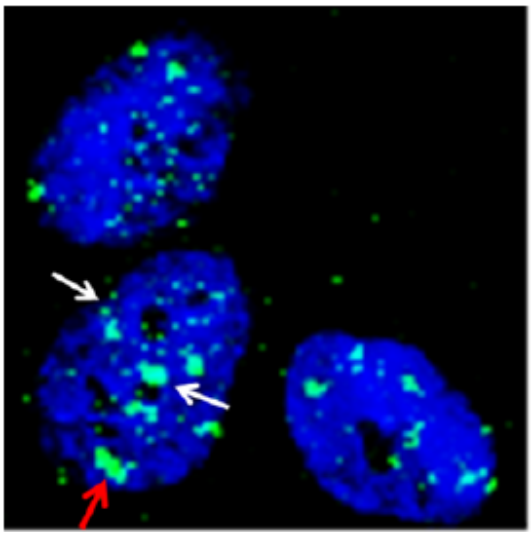

Figure 2.1: A trio of cells (blue) containing γ-H2AX foci (green), that have been stained using immunofluorescence microscopy. The two white arrows show the heterogeneity of foci sizes found in a single cell. The red arrow shows a group of foci close together that may easily be incorrectly scored as a single large focus, an example of a common scoring error [11].

2.1 Biological Background

Chromosomes in an organism are made of a substance called chromatin, which itself consists of nucleosomes in more complex, higher order structures [13]. These nucleosomes are composed of both DNA and octamers of histones, groups of eight proteins that are

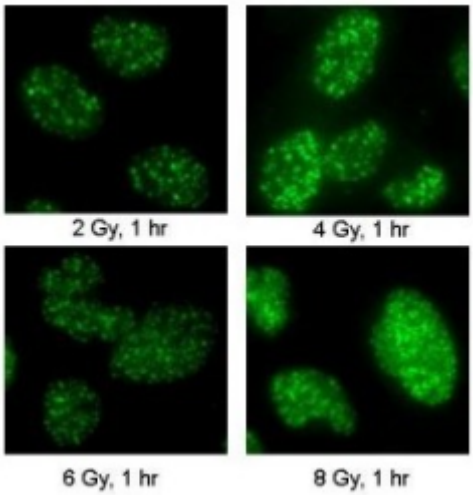

Figure 2.2: A set of four images showing the effect of increased dose on the generation of γ-H2AX foci. All samples were scored the same amount of time after irradiation. It is clear that the greater the amount of radiation exposed to, the greater the number of foci generated [12].

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 [14]. The H2A histone has four subtypes, which are grouped into three subfamilies: H2A1-H2A2, H2AZ, and H2AX [13]. We are focussing on H2AX, which can account for anything from 2% to 20% of the H2A histones found in human cells. Our data refers specifically to lymphocytes, which typically only contain 2% H2AX [15].

The H2AX subtype has a specific role, namely that it is able to repair double strand breaks (DSBs) in DNA following a nucleosome's exposure to low doses of ionising radiation [14]. When said DSBs occur, the H2AX histone phosphorylates, a process that results in the formation of γ-H2AX foci, which can be observed as fluorescent dots [6]. The background rate of γ-H2AX foci is very low (0.1 per cell or less in unirradiated peripheral blood mononuclear cells and normal human fibroblasts in stationary phase) [15]. The phosphorylation is only visible for up to approximately 24 hours after radiation exposure [6]. 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 [15].

2.2 Foci Scoring

Although the initial research on this biomarker was done using animal tissue samples, primarily mouse livers [13], this is massively impractical when dealing with irradiated humans, so blood samples are used instead.

According to Rothkamm & Horn in 2009, γ-H2AX analysis should take place between 30 minutes and 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 [15]. However, this is unrealistic when considering a radiation incident, especially a triage situation with a large number of potential casualties. Therefore, samples should be taken as soon as possible, ideally within 24 hours, with a potential upper limit for feasibility of approximately 3 days [16].

These blood samples are then scored, either manually in a process called immunofluorescence microscopy, or automatically through flow cytometry using machinery such as MetaCyte [17]. However, automated scoring can have consistency issues [18], 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 [15], as well as the discretion of the individual scorer. Indeed, this is an issue we later came across when analysing the newer of our two datasets: there were two researchers scoring the data, and one recorded consistently higher foci counts than the other. This can be accounted for by including mathematical methods to "normalise" the data and thus allow for comparison across a full dataset.

2.3 Comparison with Dicentric Biomarker

When comparing the γ-H2AX histone biomarker with its dicentric counterpart, we must first be aware that they are two different types of biomarker. The dicentric biomarker is a cytogenetic biomarker, one that counts chromosomal aberrations in blood lymphocytes, whilst the γ-H2AX histone is a protein-based biomarker that relies on the phosphorylation of proteins for analysis [6]. The dicentric biomarker is very well established in the literature, with clear and comprehensive statistical methods, whilst the γ-H2AX histone is not (the aim of our aforementioned paper [10] is to establish a unified statistical

methodology for this biomarker). A key strength of the dicentric biomarker is that it has very little inter-individual variation. This is not true for the γ-H2AX biomarker, which has, as well as potential inter-individual variation, far stronger inter-laboratory variation [14]. However, the time required between sampling and analysis is far shorter for γ-H2AX foci (a few hours) than for the dicentric biomarker (2-3 days) [18]. The dicentric biomarker also has a lower throughput than the γ-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 [7], clearly not practical for a situation with a high number of potential casualties. There is no currently stated upper limit for the number of γ-H2AX histone samples scored per week, but it is reasonable to assume that any such limit would be far larger than 3000. The γ-H2AX histone biomarker operates within comparatively strict time limits: the phosphorylated foci initially form within minutes of exposure [18], but are typically only visible until approximately 24 hours after [6]. 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 [5]. 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 [5].

Chapter 3

Constructing Dose-Response Calibration Curves

The process of dose estimation can be broken down into two steps. Firstly, a dose-response curve is fitted, which may be either linear or quadratic. The equation for this dose-response curve comes from fitting a generalised linear model (GLM) to the chosen subset of calibration data. Data is typically subsetted by time of exposure. In this chapter we will discuss the underlying statistical techniques and methodology utilised when fitting these models and thus the resulting calibration curves. The dose estimation process and the subsequent estimation of uncertainty will be discussed in Chapter 4.

3.1 Exponential Family

The exponential dispersion family (EDF) of probability distributions can be written in the following form:

$$P(z|\theta,\phi) = \exp\left[\frac{z\theta - b(\theta)}{\phi} + c(z,\theta)\right] \tag{3.1}$$

where:

- $\theta \in \mathbb{R}$ is the so-called 'natural parameter';
- $\phi > 0$ is the 'dispersion parameter', the one that we are focussing on here;
- $b : \mathbb{R} \to \mathbb{R}$ and $c : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}$ are both functions. We know that

  $$\eta = b'(\theta) \tag{3.2}$$

  is the linear predictor (see Section 3.2) and slope parameter, and

  $$\mathcal{V}(\eta) = b''(\theta) \tag{3.3}$$

is the variance function of η.

The Poisson distribution is an EDF with $z \in \mathbb{N}$. It can be expressed as

$$P(z) = \frac{e^{-\lambda}\lambda^z}{z!} = \exp\left[z \ln \lambda - \lambda - \ln z!\right] \tag{3.4}$$

for integer z [19].

In this case we have that

- $\theta = \ln \lambda$;
- $\phi = 1$;
- $b(\theta) = e^{\theta} = \lambda$;
- $c(z, \phi) = -\ln z!$.

3.1.1 Dispersion

The dispersion parameter ϕ is a parameter of the exponential family that denotes the ratio of the variance to the mean. In the Poisson case it can be calculated by dividing the variance by the mean. As the Poisson model has equidispersion, the value of ϕ for a Poisson fit is always 1. Therefore one can think of ϕ as the dispersion relative to the Poisson dispersion. If $\phi > 1$ we have overdispersion, whilst $\phi < 1$ represents (hypothetical) underdispersion.

Overall, the dispersion parameter exists to account for "extra variation" within data and is often referred to as a "nuisance" parameter, but is actually of critical importance when constructing a quasi-Linear model of any type. The size of the dispersion parameter does not affect the value of the mean, but it does affect the spread of the data, acting to scale the variance and thus the standard error.

3.2 Generalised Linear Models

Generalised Linear Models (GLMs) are a class of models that generalise linear regression where the response variable is expected to follow an EDF distribution with mean μ. They are specified by three components, namely:

- $\eta = \omega^T x$, the linear predictor [20], where ω is a p-dimensional vector of unknown parameters and x is a p-dimensional design vector of predictors that is an appropriate function of the actual covariates;

- h, an injective response function. We have that

$$\mu = E[z|x] = h(\eta) = h(\omega^T x).$$

  Its inverse $g = h^{-1}$, also written as

$$g(\mu) = \eta = \omega^T x,$$

  is the link function. This and the linear predictor form the structural assumption;

- The type of the exponential family that categorises the response function, which is the distributional assumption. In a GLM our knowledge of z given x and ω is described by an EDF whose parameters depend on x and ω:

$$P(z|x,\omega) = P(z|\theta(x,\omega),\phi) = \exp\left(\frac{z\theta - b(\theta)}{\phi} + c(z,\theta)\right).$$

  The responses Z are conditionally independent given x. Therefore:

$$P(\{z_i\}|\{x_i\}) = \prod_i P(z_i|x_i,\omega) \tag{3.5}$$

3.2.1 Poisson Models

Typically a Poisson distribution is used to model cytogenetic biodosimetry data [5]. The Poisson distribution is a count based statistical distribution with probability density function $P(z) = \frac{e^{-\lambda}\lambda^z}{z!}$ (see Equation 3.4). The single parameter λ is the mean number of times an event occurs in an interval. The Poisson distribution is notable in that both the expectation and the variance are also λ. The equality of the mean and the variance in particular is known as equidispersion [20]. However, real life data frequently violates equidispersion. With biomarker data the type of violation that we are most likely to encounter is overdispersion, where the variance exceeds the mean. (Underdispersion, where the mean exceeds the variance, can occur but is typically far rarer.)

For the aforementioned dicentric biomarker, a Poisson distribution has been shown to be sufficient for a homogeneously irradiated population of blood lymphocytes by Hilali et al in 1991 [21]. As the data we are using includes lymphocytes that have been irradiated in laboratory conditions, we assume homogeneity of irradiation. Thus, we are testing initially how well the Poisson distribution fits this γ-H2AX data, to see if the goodness of fit is dicentric biomarker specific or not, an assumption that has been made previously [22]. We are then utilising a second fitting model as a comparison, specifically the quasi-Poisson model, which will be discussed in Section 3.2.5.

Throughout this thesis, we will only be using data from whole body irradiation scenarios, i.e. ones where we assume that the entire body has been irradiated an equal amount. This is done for ease of modelling and calculation. Partial body models do exist, and they will be briefly discussed in Chapter 6.

Another issue to consider is the potential for zero inflation, which occurs when the number of zeros in a sample significantly exceeds the expected amount. This is commonly observed with count data, where the number of zeros generated by a Poisson distribution is less than the number observed [23], and is a phenomenon closely related to overdispersion. Again, this issue and potential solutions to it will be discussed further in Chapter 6.

When fitting a generalised linear model in R, a link function must be stated. The link function is a response function that relates the linear predictor to the mean function of the chosen exponential family distribution. Typically the preferred link to use statistically is the canonical link function, one that is derived from the exponential of the density function. For the Poisson distribution this is the log link $g(\mu) = \exp(\omega^T x)$. For this biomarker data, however, we have chosen to deviate from the canonical link for biological reasons, instead using the identity link $g(\mu) = \omega^T x$, as this relationship has both been empirically observed [24] and can be physically justified [15]. This guarantee is needed for feasibility, as the mean/expected Poisson count can never go below 0 for count data.

The identity link gives a fitted model of the form $\mu = \sum_i \theta_i D^i$ for a vector of coefficients $\boldsymbol{\theta}$ and powers of the dose, D, from which fitting equations can be constructed. We chose to use the coefficients A, α, and β and set $y = \mu = E(z)$, to give

$$y = A + \alpha D \tag{3.6}$$

and

$$y = A + \alpha D + \beta D^2 \tag{3.7}$$

These equations will later be used for dose estimation with a known yield y.

3.2.2 The Score Function (General Case)

A maximum likelihood estimate (MLE) with log-likelihood L satisfies the score equation $S(\hat{\omega}) = \frac{\partial L}{\partial \omega^T}(\hat{\omega}) = 0$, which comes from the score function $S(\omega) = \frac{\partial L}{\partial \omega^T}$.
In the general form we have

$$S(\omega) = \frac{\partial L}{\partial \omega^T} = \sum_i \frac{\partial L_i}{\partial \omega^T} = \sum_i \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \omega^T},$$

with

$$\frac{\partial L_i}{\partial \theta_i} = \frac{z_i - b'(\theta_i)}{\phi} = \frac{z_i - \mu_i}{\phi}.$$

Using Equations 3.2 and 3.3, we know that:

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{\partial (b')^{-1}}{\partial \mu_i} = \frac{1}{b''((b')^{-1}(\mu_i)} = \frac{1}{b''(\theta_i)} = \frac{1}{\mathcal{V}(\mu_i)}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = h'(\eta_i)$$

$$\frac{\partial \eta_i}{\partial \omega^T} = x_i$$

3.2.3 The Score Function (Poisson Case)

For this dataset, the initial type of model that we are fitting is the Poisson model with identity link $\lambda(x) = \mu(x) = \omega^T x$. The likelihood function is given by

$$L = \exp\left(\sum_i (-\mu_i) + z_i \ln(\mu_i) - \ln(z_i!)\right) = \exp\left(\sum_i (-\omega^T x_i) + z_i \ln(\omega^T x_i) - \ln(z_i!)\right)$$

We have that

$$\frac{\partial L_i}{\partial \theta_i} = z_i - \mu_i = z_i - \omega^T x_i$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\mu_i} = \frac{1}{\omega^T x_i}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = 1$$

$$\frac{\partial \eta_i}{\partial \omega^T} = x_i$$

η_i is explicitly linked to the shape of the constructed curve, as it is calculated using the curve coefficients.

Therefore in this scenario, the full score equation can be written as

$$\begin{aligned} S(\omega) &= \sum_i (z_i - \mu_i) \frac{1}{\mu_i} x_i \\ &= \sum_i \left(z_i - \omega^T x_i\right) \frac{1}{\omega^T x_i} x_i \\ &= \sum_i \left(\frac{z_i}{\omega^T x_i} - 1\right) x_i \end{aligned}$$

Here we have that $\phi = 1$ (as this is a Poisson model).

$$S(\omega) = \sum_i \left(\frac{z_i}{\omega^T x_i} - 1\right) x_i$$

3.2.4 Removing the Dispersion

$$S(\omega) = \sum_i \left(\frac{z_i - \mu_i}{\phi}\right)\left(\frac{1}{\mathcal{V}(\mu_i)}\right) h'(\eta_i) x_i$$

$$S(\omega) = \frac{1}{\phi}\sum_i (z_i - \mu_i)\frac{1}{\mathcal{V}(\mu_i)} h'(\eta_i) x_i$$

The MLE $\hat{\omega}$ must satisfy $S(\hat{\omega}) = 0$. However, the dispersion parameter ϕ cancels from the score equation. This implies that $\hat{\omega}$ does not depend on ϕ.

3.2.5 Quasi-Poisson Models

As the value of $\hat{\omega}$ does not depend on ϕ, we can construct a so-called "quasi-Poisson" model that utilises the majority of Poisson assumptions but allows for dispersion values other than 1.

In general, quasi-likelihood models drop the exponential family assumption of likelihood models. They also separate the mean and variance structure of the model, which is especially relevant in the Poisson case when the equidispersion assumption is also being removed.

Fitting a GLM using a quasi-Poisson model gives the same estimate values for coefficients as a Poisson fit, and thus the values of the residuals are the same. The null and residual deviance values are also the same. Obviously, the dispersion values differ between Poisson and quasi-Poisson fits with the same link for identical data, but so too do the standard errors on the coefficients. Quasi-Poisson models may not have distributional forms; they are only characterised by mean and variance. Thus they do not have Akaike information criteria (AIC) values. In 2002 Burnham and Anderson developed quasi-AIC, but they only used it to compare various quasi-linear models instead of comparing quasi class models to those with a distributional form (i.e. comparing quasi-Poisson to Poisson) [25].

A quasi-Poisson model is preferable to a Poisson one for our γ-H2AX data due to the amount of overdispersion present. In the examples below (Section 3.4), it can clearly be seen that the dispersion parameter for our PHE (2017) data has a value of approximately 60. This indicates a large difference in size between the mean and variance values (the variances are about 60 times larger than the means), and consequently a significant violation of the Poisson model's equidispersion principle.

3.3 Variability

3.3.1 Moments

For a GLM, the mean and variance structure are correctly specified by

$$E(z|x) = \mu = h(\omega^T x),$$

$$\text{Var}(z|x) = \sigma^2(\mu) = \phi\mathcal{V}(\mu),$$

where $\mathcal{V}(\mu)$ is a variance function [26]. It holds that the "updated" variance for the quasi-Poisson fit can be calculated using the quasi-Poisson dispersion: that is, for parameter estimates $\hat{\boldsymbol{\theta}}$,

$$\text{Var}_Q(\hat{\boldsymbol{\theta}}) = \phi\text{Var}_P(\hat{\boldsymbol{\theta}}). \tag{3.8}$$

3.3.2 Standard Error

The standard error on a measurement is the square root of the variance. Thus it follows that when moving from a Poisson model to a quasi-Poisson one we multiply the Poisson standard error SE_P by the square root of the dispersion ϕ. Therefore, for parameter estimates $\hat{\boldsymbol{\theta}}$, we have that:

$$SE_Q(\hat{\boldsymbol{\theta}}) = \sqrt{\phi}SE_P(\hat{\boldsymbol{\theta}}) \tag{3.9}$$

3.4 Examples

3.4.1 Checking Fit Type

A quick and easy way to test for equidispersion is to simply plot the means and variances of the yields against each other, with a aggregated data point for each dose. Should we have equidispersion, the resulting line of best fit will be a straight line of the form $x = y$ with the same scale on both axes. We do this in Figure 3.1, and it is immediately obvious that we have overdispersion instead of equidispersion. For starters, the means and variances are not even plotted on the same scale, with the means' scale in the thousands and the variances' scale in the hundreds of thousands. The variances are shown to be roughly 60 times larger than the means, suggesting that a fitted quasi-Poisson GLM would have an estimated dispersion of $\phi \approx 60$.

If we were to introduce a line of best fit here, it would clearly be nothing near $x = y$, and the 24h data may even be best served by fitting a curve to deal with the extremely

large value outlier on the right hand side, should it not be an outlier, although further measurements would be needed to clarify either way.

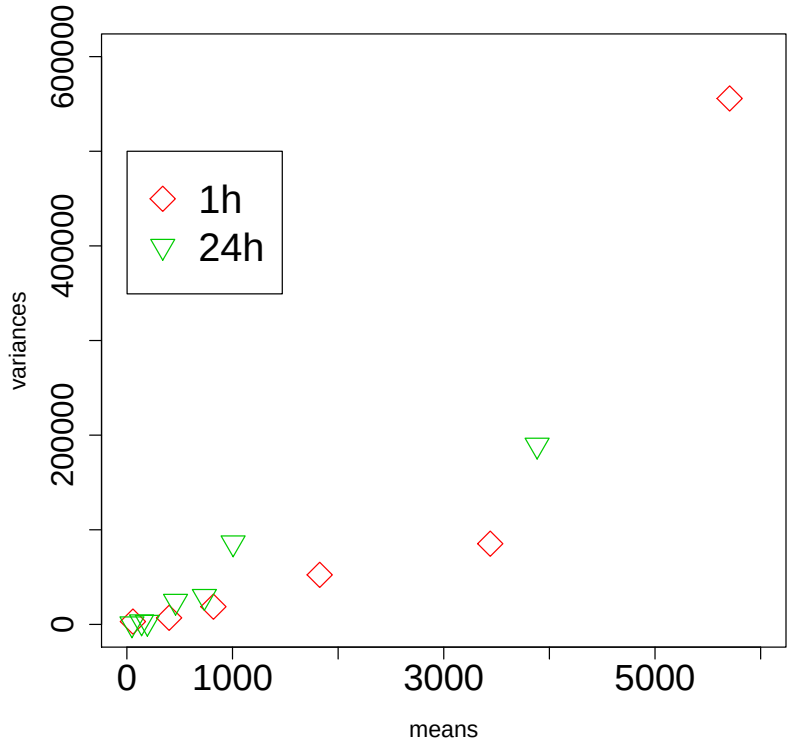

Figure 3.1: A plot showing the relationship between the means and variances of the cells in the PHE (2017) dataset. Clear differences in scale are visible, as is an obvious lack of equidispersion. From [10].

3.4.2 Model Fitting & Calibration Curves

We have a large number of data points, so we chose to fit the time points separately from each other. Initially we fitted eight different models, four Poisson and four quasi-Poisson. However, the coefficients are mathematically identical, as discussed in Section 3.2.5, thus Table 3.1 records the quasi-Poisson (rounded) coefficients, along with the dispersion values of each fit. Standard error values have been recorded separately in Table 3.2 for later discussion/analysis. These four fits have been plotted in Figure 3.2 (and colour coded appropriately) on a graph of dose against foci/cell. For quadratic fits, we expected to have that $\beta < 0$ due to the saturation effect, and indeed this is what we found.

All four of our final fits have similar intercept values of between 0.11 and 0.18, suggesting that a so-called "true" background rate lies between those values. In both the linear and the quadratic cases, the slope value decreases with time, whilst the intercept increases slightly (by about 0.03 each time), although the latter observation may not actually be notable. There is a large difference between 1h and 24h, time wise (especially compared to a gap between time points like 1h and 4h), so the large decrease (by over 10 each time) in the slope value is understandable. Both of the β coefficients are negative, as expected, with the 1h quadratic fit having a much larger curve than the 24h one. The 24h model has a very small negative curve coefficient, at only -0.094352. In fact, both quadratic fits are very close to the linear ones, according to Figure 3.2. Therefore it is especially worth looking at the uncertainty values in this case to determine whether or not the amount of uncertainty introduced by each quadratic (or β) term justifies its inclusion. It remains to be seen whether the relative sizes of the β terms appear to be "typical" for their time points, although there is a dearth of fit values to compare in this case, as linear fits are typically preferred for γ-H2AX data.

From our initial plot of means against variances, we estimated that any fitted quasi-Poisson GLMs from the data would have dispersion values of approximately 60. In the end this turned out to be a slight overestimate, but it's close enough to be sensible, especially for the linear models. The model with the largest dispersion (59.5617, just below the initial estimate) is the linear 1h fit, whilst the quadratic 1h fit has the smallest dispersion, only 43.42941. Notably, the linear fits have close together dispersion values (the difference between them is 2.01567), whilst the quadratic fits very much do not (the difference between them is 9.11545). Thus we could suggest that estimating the dispersion using a plot of means against variances is more likely to give sensible estimates for linear fits than quadratic ones. However, the introduction of the extra quadratic term does appear to reduce overdispersion, sometimes by a large amount (16.13229 in the 1h case).

3.4.3 Standard Errors of Fitted Models

Table 3.2 displays the standard errors for each fitted variable of our eight original fits, both Poisson and quasi-Poisson. These standard errors have been taken from R summary outputs for the fitted models. This allows us to use the dispersion values stated in Table 3.1 to verify the relationship between the Poisson and quasi-Poisson standard error values and thus Equation 3.9. Upon initial inspection, the Poisson standard error values are each one degree of magnitude smaller than their quasi-Poisson equivalents, which does fit.

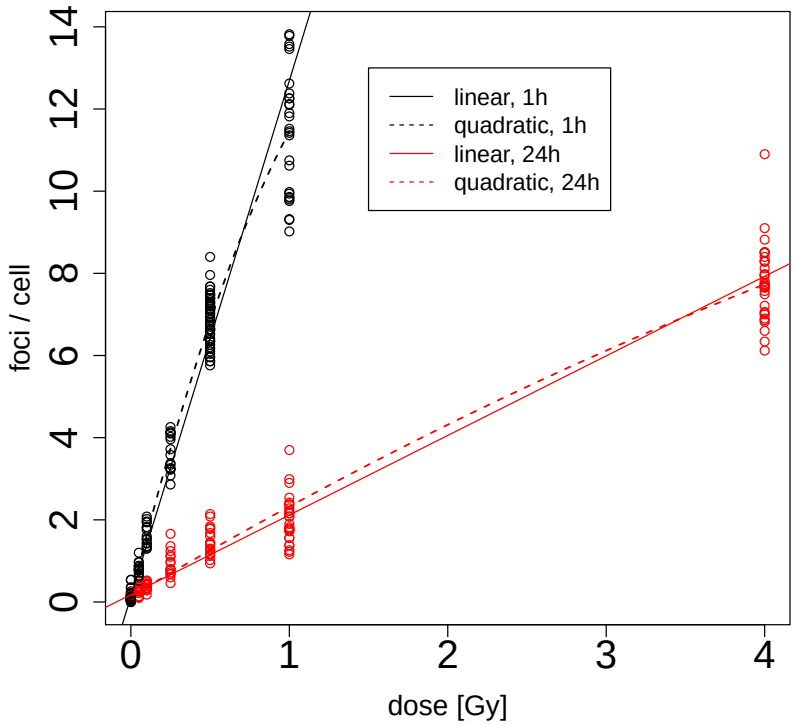

Figure 3.2: A plot showing the fitted dose-response curves for the PHE (2017) dataset, superimposed on the full dataset as shown in Figure 1.1. The linear fits are denoted by straight lines, whilst the quadratic fits are denoted by dashed ones. From [10].

Fit type	Time	A	α	β	Dispersion, ϕ
Linear	1h	0.1308	12.5589	-	59.5617
	24h	0.1794	1.9373	-	57.5460
Quadratic	1h	0.1124	15.5075	-4.1693	43.4294
	24h	0.1414	2.2756	-0.0944	52.5449

Table 3.1: A table of fitted models for the PHE (2017) dataset, showing fit type, time-point, coefficient values, and the corresponding dispersion values for the quasi-Poisson models. The model equations being fitted here are Equation 4.9 (linear) and Equation 4.10 (quadratic). Values are rounded to 4 decimal places where necessary.

Using the information given in Table 3.2, we can use Equation 3.9 to validate the dispersion values from Table 3.1, and thus verify both that our quasi-Poisson fit actually works and that Equation 3.9 is indeed what is used in R to transform a Poisson model into a quasi-Poisson model. We chose to test this using the linear and quadratic fits for

Fit type	Time	Model	$SE(A)$	$SE(\alpha)$	$SE(\beta)$
Linear	1h	Poisson	0.0021	0.0208	-
		Quasi-Poisson	0.0162	0.1606	-
	24h	Poisson	0.0028	0.0049	-
		Quasi-Poisson	0.0214	0.0375	-
Quadratic	1h	Poisson	0.0020	0.0558	0.0708
		Quasi-Poisson	0.0130	0.3675	0.4664
	24h	Poisson	0.0029	0.0151	0.0040
		Quasi-Poisson	0.0208	0.1095	0.0292

Table 3.2: A table containing the standard errors for each fitted model for the PHE (2017) dataset, sorted by fit type, timepoint, and model type, taken as they are from our R code output. Values are rounded to 4 decimal places where necessary.

the 1h data, due to them having the largest and smallest dispersion values respectively. Linear:

$$SE_Q(A) = \sqrt{59.5617} \times 0.0209 = 0.0162$$
$$SE_Q(\alpha) = \sqrt{59.5617} \times 0.0208 = 0.1606$$

Quadratic:

$$SE_Q(A) = \sqrt{43.4294} \times 0.0020 = 0.0130$$
$$SE_Q(\alpha) = \sqrt{43.4294} \times 0.0558 = 0.3675$$
$$SE_Q(\beta) = \sqrt{43.4294} \times 0.0708 = 0.4664$$

We can clearly see from these calculations that the relationship $SE_Q(\hat{\boldsymbol{\theta}}) = \sqrt{\phi} SE_P(\hat{\boldsymbol{\theta}})$ (Equation 3.9) holds, thus confirming that the quasi-Poisson functionality in R works as we would theoretically expect.

Chapter 4

Dose Estimation

4.1 Yield

In order to estimate the dose, the yield of foci from the sample must first be known. The yield, y, is found by scoring groups of cells, as discussed in Section 2.2. The number of cells scored in each sample is referred to as n. (For the PHE (2017) dataset, we have a constant n throughout. This is not the case in the PHE (2018) dataset.)

The following derivation works for the Poisson case [27]. For a single count Y_i, the Poisson distribution can be written as:

$$Y_i \sim f(\lambda, Y_i) = \frac{e^{-\lambda}\lambda_i^Y}{Y_i!}$$

for $i = (1, 2, ..., n)$. We also have that $\text{Var}(Y_i) = \lambda$ and $SD(Y_i) = \sqrt{\lambda}$. Therefore the Poisson likelihood function is given by:

$$\begin{aligned} L &= \prod_i^n \frac{e^{-\lambda}\lambda_i^Y}{Y_i!} \\ &= \frac{e^{-\lambda}\lambda^{Y_1}}{Y_1!} \times ... \times \frac{e^{-\lambda}\lambda^{Y_n}}{Y_n!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n Y_i}}{\prod_{i=1}^n Y_i!} \end{aligned}$$

and thus the log-likelihood function is:

$$\begin{aligned} ln(L) &= \sum_{i=1}^{n} ln\left(\frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} Y_i}}{\prod_{i=1}^{n} Y_i!}\right) \\ &= -n\lambda + \sum_{i=1}^{n} Y_i ln(\lambda) - \sum_{i=1}^{n} ln(Y_i!). \end{aligned}$$

We can thus find the maximum likelihood estimate (MLE) $\hat{\lambda}$ by taking the derivative of the log-likelihood with respect to λ then setting the resultant equation to 0 and solving for λ:

$$\begin{aligned} \frac{\partial ln(L)}{\partial \lambda} &= -n + \frac{\sum_{i=1}^{n} Y_i}{\lambda} \\ \Rightarrow -n + \frac{\sum_{i=1}^{n} Y_i}{\hat{\lambda}} &= 0 \\ \hat{\lambda} = \frac{\sum_{i=1}^{n} Y_i}{n} &= \bar{Y} = y. \end{aligned}$$

Thus the MLE of the count is the yield.

Knowing the MLE allows us to define the standard error (SE) of the count:

$$SE(Y_i) = \sqrt{\hat{\lambda}} = \frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n} Y_i},$$

thus we have:

$$SE(\sum_{i=1}^{n} Y_i) = \sqrt{\sum_{i=1}^{n} SE^2(Y_i)} = \sqrt{\sum_{i=1}^{n} \frac{1}{n} \sum_{i=1}^{n} Y_i} = \sqrt{\sum_{i=1}^{n} Y_i}.$$

Therefore we can define the sampling variance on the yield of foci per cell for the Poisson case as:

$$\begin{aligned} \sigma_y^2 = SE^2(y) = SE^2(\bar{Y}) = SE^2\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) &= \left(\frac{1}{n} SE\left(\sum Y_i\right)\right)^2 = \left(\frac{1}{n}\sqrt{\sum Y_i}\right)^2 \\ = \left(\frac{1}{\sqrt{n}}\frac{\sqrt{\sum Y_i}}{\sqrt{n}}\right)^2 = \left(\frac{1}{\sqrt{n}}\sqrt{\frac{\sum Y_i}{n}}\right)^2 &= \left(\frac{1}{\sqrt{n}}\sqrt{\bar{Y}}\right)^2 = \frac{\bar{Y}}{n} = \frac{y}{n} \end{aligned}$$

$$\therefore \sigma_y^2 = \frac{y}{n} \tag{4.1}$$

However, for the quasi-Poisson case we use a GLM property to transform the sampling variance. From Section 3.2.2 we know that the variance of a GLM is stated by the equation

$$\mathrm{Var}(y|x) = \sigma^2(\mu) = \phi\mathcal{V}(\mu),$$

and thus we have that, for an quasi-Poisson fit with our yield y:

$$\mathrm{Var}(y) = \phi\mathrm{Var}_P(y).$$

Therefore, the standard error on the yield for the quasi-Poisson case can be estimated using

$$\sigma_y^2 = \frac{y\phi}{n},$$

where ϕ is the dispersion parameter.

For our first dataset (PHE (2017)), the results recorded are counts, foci/cell readings for groups of $n = 500$ cells. For the second dataset, we were unable to do this due to multiple scorers (see Chapter 5) and ended up with aggregated data instead.

4.2 Dose Estimation

Now that we have a value for the yield, as well as fitted calibration curves, we can estimate the dose based on whether we have fitted a linear or quadratic model, using equations that are based on the fit of the calibration curve. For the datasets we have, linear models are preferred in the literature, but we have chosen also to give the dose estimation equation for the quadratic case to test why this is. For the linear case, we have the equation

$$D = \frac{(y - A)}{\alpha}, \tag{4.2}$$

a rearrangement of Equation 3.6 with D as the subject, whilst for the quadratic case the dose estimation equation is

$$D = \frac{-\alpha + \sqrt{\alpha^2 + 4\beta\,(y - A)}}{2\beta}. \tag{4.3}$$

This is a rearrangement of Equation 3.7 with D as the subject.

4.3 Reference Sample Ratio

In Ainsbury et al. (2016) [28], a so-called reference sample ratio, whose value is referred to as r, is included in analysis to scale the data and thus produce a suitable calibration

curve. This reference sample ratio is created by comparing yields from points on pre-existing calibration curves with "reference yields" created as controls in a lab situation. If a reference yield is slightly smaller than the yield on the curve for the same dose, we adjust by scaling the estimated doses (and thus the curve) up correspondingly. However, if the reference yield is slightly larger than the yield on the curve, no action is taken. This is because it is sensible to assume that reference yields are slightly larger as they are not subject to foci loss (such as through shipping).

The Public Health England factsheet [9] recommends that for best practice "at least one negative and one positive control" should be created for each time point in the sample. Here the phrase "negative control" refers to a non-irradiated sample, one that can be said to have been irradiated to 0Gy. We label this y_{0Gy}. The suggested dose for the "positive control" sample is 1.5Gy, so this is labelled $y_{1.5Gy}$. Creating these controls also accounts for systematic errors between laboratories. These controls should then be scored to yield pairs of control reference samples, $y_{0Gy} \pm \sigma_{y,0Gy}$ and $y_{1.5Gy} \pm \sigma_{y,1.5Gy}$. The standard error on these samples is calculated the same way as for the yield. Therefore overall, for a dataset with two time points (say 1h and 24h), one would assume that a minimum of four reference samples are needed: one irradiated to 0Gy for 1h, one irradiated to 1.5Gy for 1h, one irradiated to 0Gy for 24h, and one irradiated to 1.5Gy for 24h. However, if efficiency is required, we do not need two 0Gy control samples - a 0Gy sample serves more as a measure of background levels of foci counts and thus radiation therefore theoretically the two 0Gy samples should yield an identical foci count.

When the reference samples have been created, the 0Gy negative control values should then be compared to the corresponding fitted values of the coefficient A for the dose-response curve with the matching time point. The 1.5Gy positive control values should be compared to the 1.5Gy point on said dose-response curve. According to Public Health England [9], a discrepancy of more than 30% between the reference samples and the points on the curve is indicative of experimental conditions in the current experiment that do not "sufficiently match" the experimental conditions at the time that the calibration curves were created. An alternative method of curve validation can be done by checking whether the reference samples lie within 95% prediction intervals around our estimated calibration curves [10]. For this method, our estimated curves are validated if this is indeed the case. This method does not use the reference sampling ratio and will not be focussed on otherwise.

When we have curves validated using the first method, i.e. those where the discrepancy between the 1.5Gy points on the predefined calibration curves and the reference

samples is less than 30%, then we can use the curves to estimate dose. However, we still account for this discrepancy using a reference sample ratio for scaling.

Our next step is to create the reference sample ratio, r, using the reference sample at 1.5Gy, $y_{1.5Gy}$, and the yield on the calibration curve that corresponds to a dose of 1.5Gy, the curve point y_C. If $y_{1.5Gy} > y_C$, then $r = 1$. If $y_{1.5Gy} < y_C$, then $r = \frac{y_C}{y_{1.5Gy}} > 1$. Obviously, if $y_{1.5Gy} = y_C$, then $r = 1$. Due to the 30% discrepancy upper limit for using this method, we have that $1 \leq r \leq 1.3$. Therefore r cannot decrease the value of the estimated dose, instead keeping it the same or making it slightly larger. Ainsbury et al. (2016) [28] use a reference sample standard deviation of $\sqrt{r}$, a value that comes from Poisson assumptions. This can then be used to calculate a corresponding variance, which is needed later on when calculating the uncertainty on a dose estimate. According to [28], the standard errors on the reference samples should be estimated in the same manner as the uncertainty on the sample yield.

Now that we have this ratio r, we can incorporate it into Equations 4.2 and 4.3 as a scaling factor, giving us two further equations:

1. $$D = \frac{(y - A)}{\alpha} r \tag{4.4}$$

   (linear, comes from Equation 4.2)

2. $$D = \frac{-\alpha + \sqrt{\alpha^2 + 4\beta\,(y - A)}}{2\beta} r \tag{4.5}$$

   (quadratic, comes from Equation 4.3)

4.4 Uncertainty Analysis

With measurements comes uncertainty, and so naturally a fitted calibration curve itself has uncertainty. According to the International Atomic Energy Association (IAEA) the uncertainties associated with a dose-response calibration curve are approximately normally distributed [5].

4.4.1 Merkle's Method

When dealing with data that's been fitted to a Poisson distribution, the subsequent (Poisson) nature of the yield provides a second component to the uncertainty. It is still possible to calculate a 95% confidence interval (CI) for the "true" dose, however,

due to Merkle's method. This method is detailed in the IAEA's Dosimetry (Emergency Preparedness and Response) Manual from 2011 [5], and is well-established for use with the dicentric biomarker. It uses the confidence limits on the yield and a graph of a calibration curve. The calibration curve should have uncertainty curves plotted (preferably as dotted lines) on either side of the dose-response curve. Merkle's method includes the following steps:

1. Calculate the yields that correspond to the lower and upper 95% confidence limits on the observed yield, known as y_L and y_U respectively.

2. Calculate the dose at which y_L crosses the upper uncertainty curve. This is the lower confidence limit or D_L.

3. Calculate the dose at which y_U crosses the lower uncertainty curve. This is the upper confidence limit or D_U.

This gives you the widest possible 95% CI for this curve and yield combination.

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 attempting to shrink the interval size without first checking whether this method actually works for the γ-H2AX histone as well as the dicentric biomarker. Merkle's method works well for dicentric data because the distribution of dicentric cells can be said to be equidispersed with the Poisson distribution [29]. That is, the ratio of the sample variance to the sample mean is very close to 1:1 and, as Poisson fits are used, each model fitted with dicentric data has dispersion 1. This is not true for our γ-H2AX data, which has a far larger dispersion (around 60). Gao (2017) [29] attempted to apply Merkle's method of generating 95% confidence intervals to γ-H2AX data with known real doses. In order for the method to be successful, the expectation was that nearly 95% of the intervals produced should contain the real doses. For the 1h data, only 10 of the 131 measured yields ultimately produced intervals that encompassed the real dose (a further 4 yields gave NA as an answer). This is a success rate of 7.63%. The 24h data fared even worse with only 9 of the 130 yields producing intervals containing the real dose, a success rate of 6.92%. This is far from satisfactory and suggests that Merkle's method is insufficient for γ-H2AX data. Instead, other approaches such as the Delta Method should be considered.

4.4.2 The Delta Method

Assume we have a vector of parameter estimates $\boldsymbol{\theta}$, for which we can fully specify the variance matrix, $Var(\hat{\boldsymbol{\theta}})$. Now, assume that we are interested in a real-valued function

$h(\boldsymbol{\theta})$, and would like to find the variance of this. (This $h(\boldsymbol{\theta})$ will later correspond to the dose estimator.) From the multivariate Taylor Expansion we have that $h(\boldsymbol{\theta}) : \mathbb{R}^p \to \mathbb{R}$ for a vector $\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$. The MLE of $\boldsymbol{\theta}$ is found at the point $\hat{\boldsymbol{\theta}}$.

We know that

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \Sigma,$$

and can thus use Taylor's theorem to linearise $h(\hat{\boldsymbol{\theta}})$:

$$h(\hat{\boldsymbol{\theta}}) = h(\boldsymbol{\theta}) + \nabla h(\boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + c$$

$$\begin{aligned} \Rightarrow \mathrm{Var}(h(\hat{\boldsymbol{\theta}})) &= \nabla h(\boldsymbol{\theta})^T \mathrm{Var}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \nabla h(\boldsymbol{\theta}) + c \\ &= \nabla h(\boldsymbol{\theta})^T \mathrm{Var}(\hat{\boldsymbol{\theta}}) \nabla h(\boldsymbol{\theta}) \\ &= \nabla h(\boldsymbol{\theta})^T \Sigma \nabla h(\boldsymbol{\theta}) \end{aligned}$$

where Σ is the variance matrix.

Approximating the gradients by their estimates, we have:

$$\begin{aligned} \mathrm{Var}(h(\hat{\boldsymbol{\theta}})) &= \nabla h(\hat{\boldsymbol{\theta}})^T \Sigma \nabla h(\hat{\boldsymbol{\theta}}) \\ &= \begin{pmatrix} \frac{\partial h}{\partial \theta_1} & \cdots & \frac{\partial h}{\partial \theta_p} \end{pmatrix} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \begin{pmatrix} \Sigma_{11} & & * \\ & \ddots & \\ * & & \Sigma_{pp} \end{pmatrix} \begin{pmatrix} \frac{\partial h}{\partial \theta_1} \\ \vdots \\ \frac{\partial h}{\partial \theta_p} \end{pmatrix} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \end{aligned}$$

.

The non-diagonal terms of the variance matrix Σ correspond to covariances, written as Σ_{ij} where $i \neq j$. If $\Sigma_{ij} = 0$ for $i \neq j$ then the resulting output equation is the initial MULTIBIODOSE simplification that does not contain covariance terms:

$$\mathrm{Var}(h(\hat{\boldsymbol{\theta}})) = \begin{pmatrix} \frac{\partial h}{\partial \theta_1} & \cdots & \frac{\partial h}{\partial \theta_p} \end{pmatrix} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \begin{pmatrix} \Sigma_{11} \frac{\partial h}{\partial \theta_1} \\ \vdots \\ \Sigma_{pp} \frac{\partial h}{\partial \theta_p} \end{pmatrix} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \sum_{j=1}^{p} \left(\frac{\partial h}{\partial \theta_j} \right)^2 |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \Sigma_{jj}$$

Therefore the variance of the function is equal to the sum of the squares of the associated partial derivatives where each squared partial derivative is multiplied by the variance of the variable used to form the partial when all of the covariances are taken to be 0. We chose to do this due to the high relative uncertainty of fast biodosimetry. The covariance components have very small relative magnitude and their contributions have very little effect [28]. This choice results in a far easier equation to deal with.

The Delta Method is derived from the ISO standard (2014) and here gives a general case for estimating the uncertainty on the dose in a quadratic situation where reference samples are used. It is an immediate consequence of the derivation of the Taylor method given above.

In the case where $\hat{\boldsymbol{\theta}} = \begin{pmatrix} A \\ \alpha \\ y \\ \beta \end{pmatrix}$ and $h(\hat{\boldsymbol{\theta}}) = D$, we have that

$$\mathrm{Var}(D) = \sum_{j=1}^{4} \left(\frac{\partial D}{\partial \theta_j}\right)^2 |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \Sigma_{jj} = \left(\frac{\partial D}{\partial A}\right)^2 \Sigma_{AA} + \left(\frac{\partial D}{\partial \alpha}\right)^2 \Sigma_{\alpha\alpha} + \left(\frac{\partial D}{\partial y}\right)^2 \Sigma_{yy} + \left(\frac{\partial D}{\partial \beta}\right)^2 \Sigma_{\beta\beta}$$

$$\mathrm{Var}(j) = \Sigma_{jj} = \sigma_j^2$$

$$\Rightarrow \sigma_D^2 = \left(\frac{\partial D}{\partial A}\right)^2 \sigma_A^2 + \left(\frac{\partial D}{\partial \alpha}\right)^2 \sigma_\alpha^2 + \left(\frac{\partial D}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial D}{\partial \beta}\right)^2 \sigma_\beta^2 \tag{4.6}$$

The following equation is an immediate consequence of the general form of the Delta Method (see Equation 4.8) as derived above, in the special case $\hat{\boldsymbol{\theta}} = \begin{pmatrix} A \\ \alpha \\ \beta \\ y \\ r \end{pmatrix}$:

$$\sigma_D^2 = \left(\frac{\partial D}{\partial A}\right)^2 \sigma_A^2 + \left(\frac{\partial D}{\partial \alpha}\right)^2 \sigma_\alpha^2 + \left(\frac{\partial D}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial D}{\partial \beta}\right)^2 \sigma_\beta^2 + \left(\frac{\partial D}{\partial r}\right)^2 \sigma_r^2 \tag{4.7}$$

This equation is referred to by Ainsbury et al. (2016) as the MULTIBIODOSE simplification (specifically the case where r is included) [28]. Obviously, if reference samples are not being used we remove the corresponding variance term, resulting in Equation 4.6.

This is the general form of the Delta Method:

$$\begin{aligned} \sigma_D^2 = {} & \left(\frac{\partial D}{\partial A}\right)^2 \sigma_A^2 + \left(\frac{\partial D}{\partial \alpha}\right)^2 \sigma_\alpha^2 + \left(\frac{\partial D}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial D}{\partial \beta}\right)^2 \sigma_\beta^2 + \left(\frac{\partial D}{\partial r}\right)^2 \sigma_r^2 \\ & + 2\left(\frac{\partial D}{\partial \alpha}\right)\left(\frac{\partial D}{\partial \beta}\right) cov(\alpha, \beta) + 2\left(\frac{\partial D}{\partial \alpha}\right)\left(\frac{\partial D}{\partial A}\right) cov(\alpha, A) \\ & + 2\left(\frac{\partial D}{\partial \alpha}\right)\left(\frac{\partial D}{\partial r}\right) cov(\alpha, r) + 2\left(\frac{\partial D}{\partial A}\right)\left(\frac{\partial D}{\partial \beta}\right) cov(A, \beta) \\ & + 2\left(\frac{\partial D}{\partial A}\right)\left(\frac{\partial D}{\partial r}\right) cov(A, r) + 2\left(\frac{\partial D}{\partial \beta}\right)\left(\frac{\partial D}{\partial r}\right) cov(\beta, r) \end{aligned} \tag{4.8}$$

We prefer the MULTIBIODOSE simplification to the general form of the Delta Method due to the potential availability of the covariance values. These covariance terms would need to be derived from a previously fitted calibration curve, and may not be reported in some scenarios where we are given curves and do not fit them ourselves. We do have these values for our fitted models, but they are very small relative to the individual variances and have little impact on the uncertainty on the dose. (This may be true for all models, but requires further research.) Uncertainty analysis using the Delta Method may also be referred to as the GUM method, due to research by the Joint Committee for Guides in Metrology [30].

4.4.3 Application of the Delta Method

The uncertainty on the dose, σ_D^2, can be calculated using the variances, covariances, and partial derivatives of the variables in the dose estimation equations (A, α, y, r, as well as β for the quadratic equation).

For scenarios where a reference sample ratio is not used, the two dose equations:

1. $D = \frac{(y-A)}{\alpha}$ (linear, also Equation 4.2)
2. $D = \frac{-\alpha+\sqrt{\alpha^2+4\beta(y-A)}}{2\beta}$ (quadratic, also Equation 4.3)

have partial derivatives taken then are substituted as appropriate into the Delta Method (also Equation 4.6)

$$\sigma_D^2 = \left(\frac{\partial D}{\partial A}\right)^2 \sigma_A^2 + \left(\frac{\partial D}{\partial \alpha}\right)^2 \sigma_\alpha^2 + \left(\frac{\partial D}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial D}{\partial \beta}\right)^2 \sigma_\beta^2$$

to give:

1. $$\sigma_D^2 = \frac{1}{\alpha^2}\sigma_A^2 + \frac{1}{\alpha^2}\sigma_y^2 + \left(\frac{A-y}{\alpha^2}\right)^2 \sigma_\alpha^2 \tag{4.9}$$

   (linear)

2.

$$\begin{aligned}\sigma_D^2 &= \frac{1}{\alpha^2+4\beta(y-A)}\sigma_A^2 \\ &+ \frac{1}{\alpha^2+4\beta(y-A)}\sigma_y^2 \\ &+ \frac{1}{4\beta^2}\left(\frac{\alpha}{\sqrt{\alpha^2+4\beta(y-A)}}-1\right)^2\sigma_\alpha^2 \\ &+ \frac{1}{\beta^2}\left(\frac{(y-A)^2}{\alpha^2+4\beta(y-A)}+\frac{\alpha(y-A)}{\beta\sqrt{\alpha^2+4\beta(y-A)}}+\frac{y-A}{\beta}\right. \\ &\left.+ \frac{\alpha^2}{4\beta^2}+\frac{\alpha\sqrt{\alpha^2+4\beta(y-A)}}{2\beta^2}+\frac{\alpha^2+4\beta(y-A)}{4\beta^2}\right)^2\sigma_\beta^2\end{aligned} \tag{4.10}$$

(quadratic)

When a reference sample ratio is included, the two dose equations:

1. $D = \frac{(y-A)}{\alpha}r$ (linear, also Equation 4.4)
2. $D = \frac{-\alpha+\sqrt{\alpha^2+4\beta(y-A)}}{2\beta}r$ (quadratic, also Equation 4.5)

have partial derivatives taken then are substituted as appropriate into the Delta Method (also Equation 4.7)

$$\sigma_D^2 = \left(\frac{\partial D}{\partial A}\right)^2\sigma_A^2 + \left(\frac{\partial D}{\partial \alpha}\right)^2\sigma_\alpha^2 + \left(\frac{\partial D}{\partial y}\right)^2\sigma_y^2 + \left(\frac{\partial D}{\partial \beta}\right)^2\sigma_\beta^2 + \left(\frac{\partial D}{\partial r}\right)^2\sigma_r^2$$

to give:

1.

$$\sigma_D^2 = \frac{r^2}{\alpha^2}\sigma_A^2 + \frac{r^2}{\alpha^2}\sigma_y^2 + \frac{r^2(A-y)^2}{\alpha^4}\sigma_\alpha^2 + \frac{(y-A)^2}{\alpha^2}\sigma_r^2 \tag{4.11}$$

(linear)

2.

$$\begin{aligned}\sigma_D^2 =& \frac{1}{2\beta^2}(\alpha^2 + 2\beta(y-A) - \alpha\sqrt{\alpha^2+4\beta(y-A)})\sigma_r^2 \\ &+ \frac{r^2}{\alpha^2+4\beta(y-A)}\sigma_A^2 \\ &+ \frac{r^2}{\alpha^2+4\beta(y-A)}\sigma_y^2 \\ &+ \frac{r^2}{4\beta^2}\left(\frac{\alpha}{\sqrt{\alpha^2+4\beta(y-A)}} - 1\right)^2 \sigma_\alpha^2 \\ &+ \frac{r^2}{\beta^2}\left(\frac{(y-A)^2}{\alpha^2+4\beta(y-A)} + \frac{\alpha(y-A)}{\beta\sqrt{\alpha^2+4\beta(y-A)}} + \frac{y-A}{\beta}\right. \\ &\left.+ \frac{\alpha^2}{4\beta^2} + \frac{\alpha\sqrt{\alpha^2+4\beta(y-A)}}{2\beta^2} + \frac{\alpha^2+4\beta(y-A)}{4\beta^2}\right)^2 \sigma_\beta^2\end{aligned} \tag{4.12}$$

(quadratic)

In order to calculate these uncertainty values (for a given input dose, model, foci count, and number of cells scored at that dose), we used R code. This is the code for the simplest scenario, the linear situation without a reference sample ratio/where the reference sample ratio is taken to be 1:

```
s2D.li<- function(y,fit,n,count){
 A1<- fit$coefficients[[1]]
 a1<- fit$coefficients[[2]]
 d1li<-summary(fit)$dispersion
 s2A1<- vcov(fit)[1,1]
 s2a1<- vcov(fit)[2,2]
 s2y1<- d1li*y/n
 q1<-s2A1/a1^2
 q2<-s2y1/a1^2
 q3<-(A1-y)^2*s2a1*a1^-4
 q=c(q1,q2,q3)
 names(q)=c("A","y","alpha")
 barplot(q, xlab=expression(paste("Coefficients of ", sigma^2)),
 ylab="Value",main="Linear")
 sigma2D1<- q1+q2+q3
 print(q)
 return(sigma2D1)
```

```
}
```

Here $A1$ is the intercept coefficient A and $s2A1$ its variance σ_A^2, whilst $a1$ is the slope coefficient α with variance σ_α^2 inputted as $s2a1$. $s2y1$ is the variance σ_y^2 of the yield. We are implementing Equation 4.11 here.

The full R scripts for the other scenarios can be found in Appendix C.

We decided to split the calculations up into what we chose to refer to as "terms": the coefficients of each variable's variance σ^2 multiplied by the variance itself, which allows for easier error checking (and reduces the amount of calculations done at a time in R). The A, α, and β terms are referred to as "coefficient" terms, as the value of the labelling coefficient is generated by the fitted GLM. For instance, $\frac{1}{\alpha^2+4\beta(y-A)}\sigma_A^2$ is the A term from the quadratic Delta Method equation that doesn't include a reference sample ratio (Equation 4.10). In this case, σ_A^2 is the variance of A, the intercept variable, and $\frac{1}{\alpha^2+4\beta(y-A)}$ is its associated coefficient. Our R formula then sums the term values to get an output variance value.

We also realised that the linear expression without reference sample ratio could be reformulated into a quadratic equation with terms of y:

$$\sigma_D^2 = \frac{\sigma_A^2}{\alpha^4}y^2 - \frac{2A\sigma_\alpha^2}{\alpha^4}y + \frac{1}{\alpha^2}\left(\sigma_A^2 + \sigma_y^2 + \frac{A^2}{\alpha^2}\sigma_\alpha^2\right) \tag{4.13}$$

then coded this separately in a similar manner. We chose to label this the "y reformulation".

Upon seeing (and subsequently being surprised by) the (unexpectedly large) uncertainty values that the code produced for our original dataset, we chose to graph the terms, plotting each one separately on a bar chart. This allowed us to see which ones contributed most or least to the overall uncertainty for each dose. If one coefficient term contributes significantly more to the uncertainty than all of the others, this implies that the preferable fitted model is one that does not contain that term. As shown later in Figure 4.1, the β (or quadratic) term gives a significantly larger uncertainty than the other terms, so we should prefer a linear fitted GLM over a quadratic one.

4.5 Examples

4.5.1 Dose Estimation

Now that we have the coefficients of our fitted model, we can use Equations 4.2 and 4.3, as specified earlier to estimate the dose of exposure for a chosen yield. We initially

choose to do this four times (linear and quadratic scenarios for both timepoints) to give us four comparative doses. We can then double check our calculated doses against the calibration curves. Our four dose equations are as follows:

1h:

Linear:

$$D = \frac{(y - 0.1308)}{12.5589}$$

Quadratic:

$$\begin{aligned} D &= \frac{-15.5075 + \sqrt{15.5075^2 + 4 \times -4.1693\,(y - 0.1124)}}{2 \times -4.1693} \\ &= \frac{-15.5075 + \sqrt{15.5075^2 - 16.6773\,(y - 0.1124)}}{-8.3386} \end{aligned}$$

24h:

Linear:

$$D = \frac{(y - 0.1794)}{1.9373}$$

Quadratic:

$$\begin{aligned} D &= \frac{-2.2756 + \sqrt{2.2756^2 + 4 \times -0.0943\,(y - 0.1414)}}{2 \times -0.0944} \\ &= \frac{-2.2756 + \sqrt{2.2756^2 - 0.3774\,(y - 0.1414)}}{-0.1887} \end{aligned}$$

For an arbitrarily chosen yield of 5 foci/cell, $y = 5$, we can thus estimate the doses for all four scenarios. This yield was chosen as it is large enough to give very different dose estimates for the two different timepoints, as well as hopefully providing differing dose estimates between the pairs of linear and quadratic fits. All doses stated here are rounded to four decimal places.

1h:

Linear:

$$D = \frac{5 - 0.1308}{12.5589} = 0.3877Gy$$

Quadratic:

$$\begin{aligned} D &= \frac{-15.5075 + \sqrt{15.5075^2 - 16.6773\,(5 - 0.1124)}}{-8.3386} \\ &= \frac{-15.5075 + \sqrt{158.9705}}{-8.3386} \end{aligned}$$

$$D = 0.3477Gy$$

24h:

Linear:

$$D = \frac{5 - 0.1794}{1.9373} = 2.4884Gy$$

Quadratic:

$$\begin{aligned} D &= \frac{-2.2756 + \sqrt{2.2756^2 + 4 \times -0.0944\,(5 - 0.1414)}}{2 \times -0.0944} \\ &= \frac{-2.2756 + \sqrt{3.3445}}{-0.1887} \end{aligned}$$

$$D = 2.3675Gy$$

As expected, the estimated doses for the 24h scenarios are significantly larger than those for the 1h scenarios, with a difference of over 2Gy between timepoints for each model type.

4.5.2 Uncertainty Quantification

Initially we had twelve scenarios, six Poisson and six quasi-Poisson. The scenarios we used were for both timepoints and were the standard linear and quadratic fits (Equations 4.9 and 4.10 respectively), as well as the aforementioned "y reformulation" (Equation 4.13) above. We chose not to include the reference sampling ratio at this time, but did test these scenarios with two different yields (5 foci/cell and 1 foci/cell). To decide which scenarios were most beneficial, we chose to code and plot all of them (Figures D.1 and D.3 in the Appendices respectively) and then analyse those plots. Unsurprisingly, the y reformulation had far more uncertainty attached to the "constant" term than either the y^2 or y coefficients, which is to be expected due to the greater number of terms involved (three, compared to one each in the other two cases). Upon further inspection of those twelve initial plots, we decided to prioritise a set of four plots and their corresponding uncertainty equation scenarios. These four scenarios are the standard linear and quadratic quasi-Poisson fits for both the 1h and 24h timepoints, and are shown later in Figures 4.1 (yield of 5 foci/cell) and D.2 (yield of 1 foci/cell). Ultimately, we decided that although interesting, the y reformulation is of less immediate interest compared to the more "conventional" way of splitting up the uncertainty equations. The quasi-Poisson fits are of more interest to us than the Poisson ones as the total variance for a quasi-Poisson

fit will be significantly larger than the Poisson total variance due to Equation 3.8. This equation can also be applied to each parameter in the model separately, therefore the relative size of the bars on a corresponding pair of plots (one Poisson and one quasi-Poisson) will be the same. Using the plots with larger values allows us to see just how big the variances for a certain yield (and thus dose estimate) could get.

As in Section 4.5.1, we use a sample yield of 5 foci/cell to begin with. (A second yield of 1 foci/cell is shown in Appendix D, along with the full set of plots for all 12 scenarios for the yield of 5 foci/cell.) We used the R code given in Section 4.4.3 and Appendix C to calculate the values for each "coefficient" and ultimately create the plots shown in Figure 4.1.

Fit type	Time	Coefficients of σ^2				Total uncertainty	Standard error
		A	y	α	β		
Linear	1h	1.6565×10^{-6}	3.7763×10^{-3}	2.4581×10^{-5}	-	3.8025×10^{-3}	6.1665×10^{-2}
	24h	1.2197×10^{-4}	1.5334×10^{1}	2.3196×10^{-3}	-	1.5578×10^{-1}	3.9469×10^{-1}
Quadratic	1h	1.0646×10^{-6}	2.7319×10^{-3}	1.0270×10^{-4}	1.1143×10^{-1}	1.1427×10^{-1}	3.3803×10^{-1}
	24h	1.2947×10^{-4}	1.5711×10^{-1}	2.0101×10^{-2}	3.4800×10^{1}	3.5077×10^{1}	5.9226×10^{0}

Table 4.1: The individual "term" values for Figure 4.1, as well as their sum, the total uncertainty for the estimated dose. All values given in standard form, to five significant figures.

It is clear from Table 4.1 that overall the quadratic fits have significantly larger total uncertainties than the linear ones, typically by two orders of magnitude. The uncertainty associated with the β term provides the bulk of these larger uncertainty values, and, in the case of the 24h quadratic fit, has a value of over 30. If we construct confidence intervals of ± 2 standard errors for the doses estimated from our quadratic fits, we end up with 0.3476761 ± 0.6760696 for the 1h fit, and 2.367536 ± 11.84517 for the 24h fit. Clearly, these are not possible given that the minimum possible dose for exposure is 0Gy. As well as this, the width of the 24h quadratic fit interval is so large as to be effectively useless, considering that the maximum dose used in the dataset is 4Gy, due to how hard it is to count foci beyond that point. The size of the uncertainties associated with the β terms provides evidence to reject quadratic fits in favour of linear ones for this biomarker.

4.5.3 Reference Sample Ratio

We already know that if $y_{1.5Gy} > y_C$, then $r = 1$. If $y_{1.5Gy} < y_C$, then $r = \frac{y_C}{y_{1.5Gy}} > 1$. Obviously, if $y_{1.5Gy} = y_C$, then $r = 1$.

For this example, we are using raw data from our second γ-H2AX histone dataset (PHE (2018)) as reference samples. This dataset is the focus of Chapter 5, but little knowledge of it is required here as we are only using the four pieces of sample data listed in Table 4.2. One quirk that must be noted is that since we are only using two reference samples per time point any quadratic curves that are rejected will be replaced with linear ones.

Dose (Gy)	Time (h)	Yield (foci/cell)	30% Discrepancy Interval
0	1	0.34	(0.238, 0.442)
	24	0.585	(0.4095, 0.7605)
1.5	1	5.02	(3.514, 6.526)
	24	2.915	(2.0405, 3.7895)

Table 4.2: A table of the raw data from the PHE (2018) dataset used as reference samples (y_{0Gy} for the 0Gy yields and $y_{1.5Gy}$ for the 1.5Gy yields) in Section 4.5.3, along with the interval that gives a 30% discrepancy limit on each side.

Dose (Gy)	Time (h)	Fit type	Yield (foci/cell)
0	1	Linear	0.1308
		Quadratic	0.1124
	24	Linear	0.1794
		Quadratic	0.1414
1.5	1	Linear	18.9691
		Quadratic	13.9927
	24	Linear	3.0853
		Quadratic	3.3424

Table 4.3: A table of the calculated calibration yield values (y_C) used in the reference sampling ratio examples in Section 4.5.3, rounded to 4 decimal places where necessary.

Initially we used all four of our calibration curves to calculate yield values at the chosen control dose points of 0Gy and 1.5Gy, as shown in Table 4.3. These are our calibration yields. We then cross-referenced these with the discrepancy intervals given in Table 4.2, and it is apparent from doing so that only the two 24h, 1.5Gy fitted values are less than 30% above or below (in both cases, above) the reference yield values. This means that none of the curves can be validated fully overall. However, one can argue that the positive control sample is of greater importance than the negative (0Gy) one. Both

the value of A and its 30% discrepancy interval are comparatively small, and it would be undesirable to throw a curve away due to that mismatch [10]. Therefore we ultimately chose to calculate a pair of reference sampling ratios for our two "validated" yields. In both of these cases our sample yield is greater than the reference yield, so

$$r = \frac{y_C}{y_{1.5Gy}}.$$

For the 24h linear fit at a dose of 1.5Gy:

$$r_{24L} = \frac{3.0853}{2.915} = 1.0584.$$

For the 24h quadratic fit at a dose of 1.5Gy:

$$r_{24Q} = \frac{3.3424}{2.915} = 1.1466.$$

Using these values we would then be able to adjust the dose equations and calculate dose estimates (rounded to 4 decimal places) for a yield of 5 foci/cell:

Linear:

$$\begin{aligned} D_{24L} &= \frac{(y - 0.1794)}{1.9373} r_{24L} \\ &= \frac{(y - 0.1794)}{1.9373} 1.0584 \\ &= 0.5463\,(y - 0.1794) \end{aligned}$$

$$D_{24L} = 2.6337Gy$$

Quadratic:

$$\begin{aligned} D_{24Q} &= \frac{-2.2756 + \sqrt{2.2756^2 + 4 \times -0.0944\,(y - 0.1414)}}{-0.1887} r_{24Q} \\ &= \frac{-2.2756 + \sqrt{2.2756^2 + 4 \times -0.0944\,(y - 0.1414)}}{-0.1887} 1.1466 \\ &= -6.0764 \left(-2.2756 + \sqrt{2.2756^2 + 4 \times -0.0944\,(y - 0.1414)}\right) \end{aligned}$$

$$D_{24Q} = 2.7147Gy$$

Both of these estimates are larger than those from the other 24h equations in Section 4.5.1, with a larger difference between the original quadratic dose estimate and the adjusted one. This can be explained by the larger reference sampling ratio value, and shows that greater upward scaling is needed for the quadratic fit than for the linear one.

Due to the negative control mismatch, we could alternatively construct replacement calibration curves using the reference yields. These curves, constructed from 2 points, are $y = 3.12D + 0.34$ for the 1h data and $y = 1.553333D + 0.585$ for the 24h data. With regards to the uncertainty, the relative proportions of the existing terms would remain the same, as each one is multiplied through by r^2.

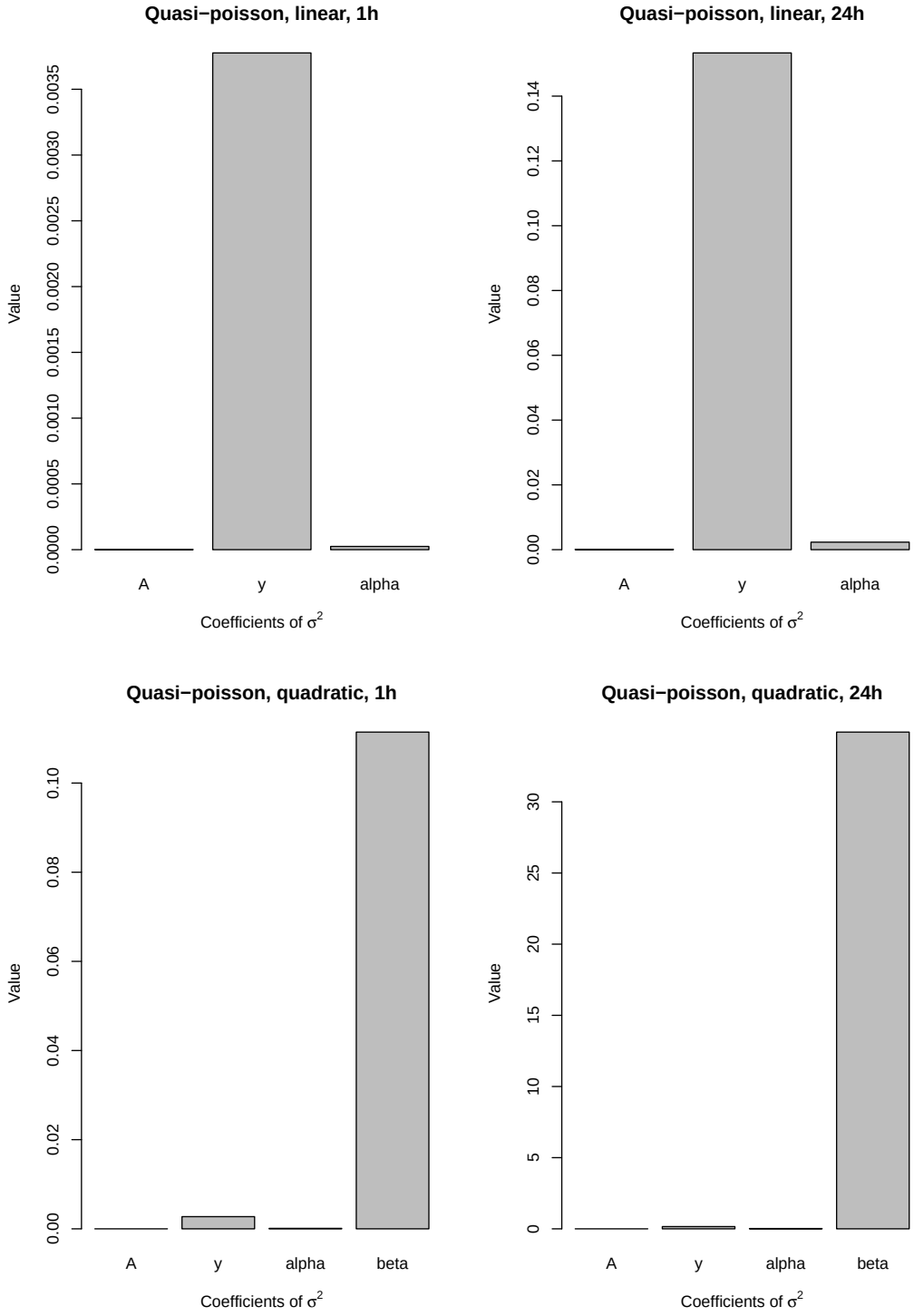

Figure 4.1: A series of four plots showing the contribution to the overall uncertainty for each "term" in a given scenario with a yield of 5 foci/cell, accompanying the code given in Section 4.5.2 and Appendix C. The four scenarios included are all quasi-Poisson models. The top row shows term by term uncertainty for linear scenarios at both timepoints (1h and 24h), whilst the bottom row shows the pair of corresponding quadratic fits.

Chapter 5

Introducing a Second Dataset

5.1 Background

The second dataset that we have worked with was also provided by Public Health England and uses blood lymphocytes from a single donor, whose blood was taken at three different times. It will henceforth be referred to as the PHE (2018) dataset. Irradiation was performed in vitro, with 250kVp X-rays as the radioactive source, and the slides were then manually scored. This dataset contains measured foci counts for four design doses: 0.75Gy, 1.5Gy, 3Gy, and 4Gy, as well as a control dose, 0Gy. All of these doses had samples scored at three different time points: 1h, 4h, and 24h. The initial dataset consists of both full body and simulated partial body scored samples, at 10% intervals, but as we are focussing on full body irradiation we chose to separate out the 0% and 100% irradiation data for use in model fitting. We later decided to code this in R using an indicator variable whose value is 0 if the sample hasn't been exposed to radiation or 1 if it has been. The amount of 0Gy data is significantly larger than the amount of data for any other dose, because the experiment was set up so that every dose & time point combination has two sets of data, one irradiated and one not.

The intended number of cells scored for each dose & time point combination was 200. However, there is some data missing due to equipment (specifically slide) issues when scoring. The 4Gy data is especially affected by this, with no irradiated 1h or 24h data remaining. One reading is missing for the control data at the 24h time point within the 1.5Gy sample. There are also issues with the 3Gy data: 4 readings are missing for the irradiated 4h data, whilst 147 readings are missing for the irradiated 1h data, nearly three-quarters of the scored cells for that dose & time point combination. We thus have potential issues with the usefulness of the 0.75Gy irradiated 1h data due to the

significantly reduced sample size, and the 4Gy irradiated data overall only exists for one of the 3 timepoints, reducing the availability of comparison.

5.2 Work Done

5.2.1 Scorers

From initial information given, we know that there are two scorers, one of whom routinely recorded higher foci counts than the other at a dose of 0Gy. The raw data labels these "Scorer 1" and "Scorer 2". Scorer 1 was responsible for 16 of the 22 sets of readings, whilst Scorer 2 was responsible for the other 6 sets. We chose to plot each scorer's measured foci counts for a dose of 0Gy, and from these plots (Figure 5.1) we can see that there is a clear difference between the counts (and thus the yields) that they recorded. Scorer 2's maximum count is only 3 foci, recorded 3 times, whilst Scorer 1 not only records a count of 3 foci many more than 3 times - they also have a maximum recorded count of 7 foci. Knowing which scorer records higher foci counts makes it easier for us to normalise the data to account for the differences between the two of them.

After observing in Figure 5.1 that there appeared to be a difference between scorers, we then wanted to determine whether or not this was large enough to be statistically significant. To do this we performed a Welch's two sample t test (which does not assume equal variances) using the 0Gy data. Our null hypothesis is that the population means for the two groups are equal, or $\mu_1 - \mu_2 = 0$. This gave us a t value of 16.123 on 2245.2 degrees of freedom, with a stated p value of $p < 2.2 \times 10^{-16}$. As this p value is less than 0.05, we reject this null hypothesis. A 95% confidence interval for the difference between the group means $\mu_1 - \mu_2$ was given as $(0.3994089, 0.5100251)$, and the difference between the sample estimates was $0.547217 - 0.092500 = 0.454717$. Therefore we have a statistically significant difference between the two sets of data, which needs to be accounted for in order to successfully fit models for the whole dataset.

5.2.2 Data Cleaning

To correct for the differences between the scorers, PHE scientists applied some initial normalisation to the data, relative to the 0Gy readings and accounting for differences between scorers. A screenshot of some of this data is given in Appendix A, specifically in Figure A.2. When analysing the data we initially calculated the mean value of the normalised 0Gy data. However, this turned out to be -0.0001709045, which is less than

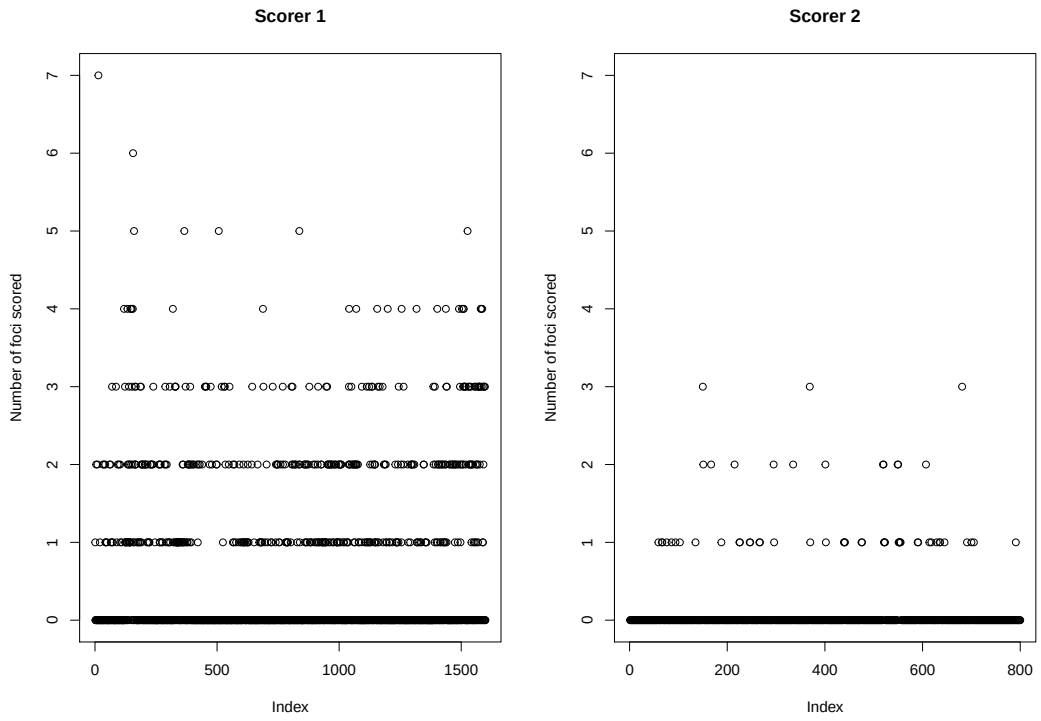

Figure 5.1: A pair of plots with equal scales showing the foci counts obtained by each scorer at a dose of 0Gy, sorted by index number in the PHE (2018) dataset. The data that will be used for model fitting is an aggregated form of this.

0. This presents an issue for fitting both Poisson and quasi-Poisson models, as mean values below zero violate the modelling assumptions.

To correct for the many negative values in the normalised original data, especially the 0Gy data, we calculated the mean of all the original scored counts for the 0Gy values (note: *not* the normalised values) and summed it with each normalised value. The resulting set of (hundreds of) values are known as the "corrected" ones. We then aggregated the data by taking the mean of the corrected values for each cell, resulting in 22 data points. However, 12 of those data points correspond to a dose of 0Gy, and it turned out that all 12 had the same corrected value of 0.3955815, meaning that only 11 distinct data points will be visible on any plotted graph.

5.3 Examples

5.3.1 Model Fitting

There are some potential issues when fitting a GLM to this data. For a given dose, we have at most 8 data points, 4 of which have the same value. Thus we only have a maximum of 5 distinct values to fit dose-response curves with, which is very low. Our initial attempts to fit models suffered as a result of this, and the presence of underdispersion was a useful indicator of our errors.

Instead we ultimately decided to try and jointly fit the time points together. In this model we took 1h as the "default" exposure time and used terms containing indicator functions to produce variables that also applied only when the time points were 4h or 24h for both the intercept and the slope coefficients. We also changed the response variable from the "corrected" values to their product with n, the total number of readings in each cell. We labelled this new variable "ncorr".

The R code from our fitted model is as follows:

```
fit.ndt.li.quasi<-glm(ncorr~-1 + n + ndose.actual + ntime4+ntime24
+ntime4dose.actual+ntime24dose.actual,
family="quasipoisson"(link="identity"), data=pheh2axc)
fit.ndt.li.quasi
fit.ndt.li.quasi$coefficients
#n                 ndose.actual        ntime4            ntime24
#0.43288416         2.53297659         0.07983618        -0.03947914
#ntime4dose.actual ntime24dose.actual
#-1.26913989        -1.49237194
summary(fit.ndt.li.quasi)$dispersion
#[1] 49.73107
```

This new dispersion value is far closer to the expected 60 and shows evidence of overdispersion, as expected.

We can split our new fitted model into three linear models, one for each timepoint, by summing the relevant terms, as shown in Table 5.1. These fits are also plotted below as fit lines on a graph of the aggregated data.

To verify that fitting the time points together was a suitable idea, we also tested it with our earlier dataset. The resultant linear fits produced were very similar to our established ones, suggesting that this method is worth using for another dataset with fewer aggregated data points.

Time	Intercept, A	Slope, α
1h	0.43288416	2.53297659
4h	0.5127203	1.263837
24h	0.393405	1.040605

Table 5.1: A table of the coefficient values for the three linear dose-response curves calculated from the joint quasi-Poisson model fitted using the PHE (2018) dataset.

Of our 3 fits, the 4h one has the highest intercept value, whilst the 1h fit has the greatest slope. These slope values follow the general established pattern: the greater the amount of time since the sample's exposure to radiation, the lower the value of the slope coefficient. Notably, the difference between the 1h and the 4h slope coefficient values is far larger than the difference between the 4h and the 24h values.

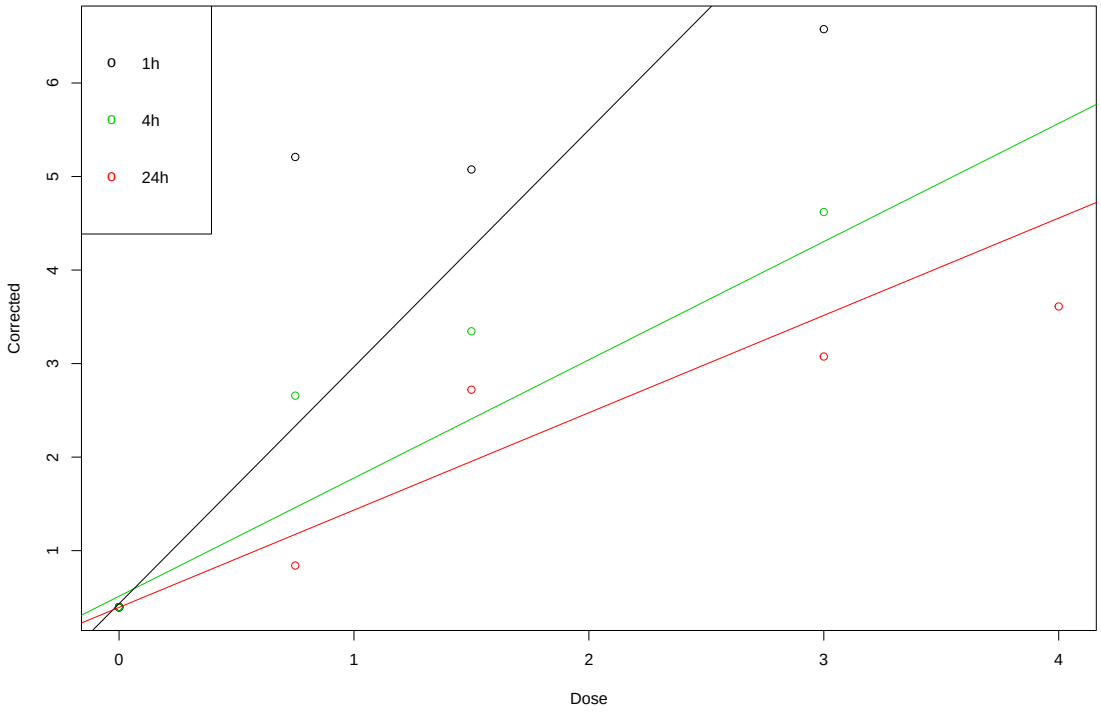

Figure 5.2: A plot showing the fitted dose response curves for the aggregated data from the PHE (2018) dataset, along with the aggregated data itself. The 1h line and data points are shown in black, the 4h in green, and the 24h in red.

Data	Time	Intercept, A	Slope, α	Notes
Public Health England (2018)	1h	0.4329±0.1633	2.5330±0.3884	
	4h	0.5127±0.1734	1.2638±0.2111	
	24h	0.3934±0.1533	1.0406±0.2572	
Public Health England (2017)	1h	0.1308±0.0162	12.5589±0.1606	
	24h	0.1794±0.0214	1.9372±0.0375	
Rothkamm et al. (2013)	4h	0.10±0.09	1.47±0.22	1L
		0.12±0.07	3.10±0.21	2L
		0.71±0.19	1.08±0.17	5L
		0.35±0.26	1.48±0.26	5B
	24h	0.08±0.02	0.54±0.04	1L
		0.09±0.02	0.78±0.03	2L
		0.46±0.09	0.94±0.09	5L
		0.13±0.03	0.70±0.05	5B
Ainsbury et al. (2016)	4h	0.6454±0.0822	2.4686±0.0676	HDR
	24h	0.1060±0.0379	0.8227±0.0455	HDR

Table 5.2: Table of linear fits for comparison purposes including the standard errors for each term. For the Rothkamm et al. (2013) data the number given in the notes refers to the number of the lab that the curve was calibrated at. The letter refers to the sample type, either B for a (full) blood sample or L for lymphocytes. The Ainsbury et al. (2016) data in the bottom two rows of the table is the data for the γ-H2AX sample recorded in that paper, which is itself cited as being from Rothkamm et al. (2013).

5.3.2 Comparing Fitted Models

When selecting which fits to use from the data presented by Rothkamm et al. (2013) [7] we chose only to use those where the cells had been scored manually, to match the manual scoring performed for the PHE datasets. We are also only comparing linear fits here, due to the greater number available for comparison.

Aside from our new fitted dataset, we only have one other 1h dataset for comparison (our other PHE dataset), and the old dataset has an intercept roughly a third of the height of the new. The slope of the old is approximately 5 times greater than that of the new. Overall, there is very little to compare here. This lack of samples may be due to the relative impracticality of a 1h time point, as in a real life situation it would be unrealistic to collect and score a sample from a patient within this time frame.

At first glance, the coefficients for our new 4h fit seem to make sense compared to

the pre-existing ones. The intercept of our new fit is high compared to labs 1 and 2, but it is sufficiently close in value to the other data (within 0.2 either side) to not feel out of place. It is within 1 standard error for 3 out of the 6 models it is being compared to. However, the 4h slope for our new dataset is the second smallest slope value for that timepoint, although it is within 1 standard error for the lab 1 model and the lab 5 blood sample model, so there isn't too much cause for concern.

When it comes to our new 24h fit, the intercept value is unexpectedly high. It is closer to the lab 5 lymphocyte data, which some may suspect to be anomalous, than any of the other 5 intercept values. Our intercept value is similar to the one we fitted for 1h data though, which could suggest a relatively constant background rate throughout. It is still over twice the size of the smaller 5 intercept values. On the other hand, our new slope value is just over half that of the old 24h PHE data and just over the upper interval limit for lab 5 lymphocytes. Compared to all other fits our new slope value is large but not significantly so.

We know that γ-H2AX dose-response curves are strongly time dependent, and have now seen these trends across datasets, so how would we deal with data where the time since exposure is unknown? Yield can be used to calculate a dose and the uncertainty on that dose, but here we run into a practical issue: we don't know if a low foci yield comes from a low dose or a high dose but with a long period of time between radiation and sampling. Horn, Barnard, and Rothkamm (2011) address modelling foci decay over time [22], and equations like those they suggest could be of use, but further research is needed to estimate decay time. Here again for unknown time and dose the primary complication remains that their effects are interlinked and difficult to separate.

Chapter 6

Further Topics & Conclusions

6.1 Further Topics

In this section we will address other radiation dose modelling topics that did not form part of our earlier modelling and analysis but are still of interest.

6.1.1 Partial Body Exposure

It is easiest statistically to assume that an entire body has been homogeneously exposed to ionising radiation. However, this does not reflect real life: according to Horn, Barnard, and Rothkamm (2011) [22], "most human radiation exposures are partial body exposures". Partial body exposure occurs when a dose of ionising radiation is received inhomogeneously across a person. Not all of the person may have been exposed. Using a whole body dose estimate for these partial body exposures can prove dangerous, due to both the underestimation of the peak dose delivered to a body part and the potential for this resulting in incorrect triage or treatment for the patient [22]. For instance, highly localised exposure to large doses of ionising radiation may result in severe tissue damage. Early clinical intervention could potentially reduce or even prevent this, so knowing the magnitude of the peak dose delivered could have a significant impact on patient health. Typically, partial body estimation is simulated in lab conditions by mixing two blood samples, one irradiated and one not, in stated and easily controlled proportions (e.g. 40% irradiated and 60% not). These two blood samples do not have to be from the same donor - Hilali et al. (1991) irradiated a blood sample from a male donor and mixed it with a not irradiated blood sample from a female one, later using the Y chromosome from the male donor as an indicator [21].

For the dicentric assay, which is typically modelled using a Poisson distribution for

whole body exposure, overdispersion is a hallmark of partial body exposure. Whole body exposure to sparsely ionising radiation leads to aberrations being induced randomly in all cells, whilst a partial body exposure results in both a high number of cells without aberrations and those with multiple dicentrics [22]. This results in overdispersion for scored cells, which can be used to estimate the extent of the irradiation [4]. We assume that a given blood sample consists of two portions, or fractions: a fraction of Poisson distributed dicentrics from the irradiated part, and a fraction of dicentric-free cells from the remaining unirradiated part of the body. (Note: this assumes that the irradiation of the exposed section is homogeneous.) An iterative maximum likelihood method can then be used to approximate the fraction, f, of the scored cells that were irradiated and the corresponding mean yield, y (Hilali et al. refer to this y as the "aberration frequency in [the] population" [21]):

$$\frac{y}{1-e^{-y}} = \frac{X}{N-N_0} \tag{6.1}$$

$$f = \frac{X}{Ny},$$

where N is the total number of cells scored, X is the number of dicentrics observed, and N_0 is the number of cells without dicentrics. We can only solve Equation 6.1 iteratively. Standard dose-response curves can then be used to estimate the mean dose to the irradiated fraction of the body [4]. However, we cannot derive the actual fraction of the body exposed to radiation without correcting for the effects of mitotic delay and apoptosis (interphase death), the factors which can reduce cells' ability to reach metaphase in a 48h culture. This reduction of the observable irradiated fraction is dose-dependent and can be calculated using:

$$P = e^{\frac{-D}{D_0}},$$

where P is the surviving fraction, D is the absorbed dose, and D_0 is the dose required to reduce the number of irradiated cells to 37% of the original due to the factors mentioned above, assuming an exponential dose-effect relationship. The corrected fraction of irradiated cells, F, can thus be calculated using:

$$F = \frac{f/P}{1-f+f/P}.$$

This is known as the contaminated Poisson method.

An alternative available method for dose estimation in a partial body exposure situation is the *Qdr* method. It uses the yield of aberrations (dicentrics and rings) in just

the damaged cells to approximate the dose delivered in the exposure, assuming that these aberrations are indeed caused by the exposure to ionising radiation and thus were present when it took place. The expected yield, labelled Qdr, of aberrations amongst the damaged cells, N_d, is given by:

$$Qdr = \frac{X}{N-d} = \frac{y_1}{1-e^{-y_1-y_2}},$$

where y_1 and y_2 are yields of dicentrics plus rings and excess acentric aberrations, respectively [4]. The limitations of this method include the lack of information on the size of the irradiated fraction, as well as two incorrect assumptions: that the background frequency of dicentrics and rings is 0, and that excess acentrics fit a Poisson distribution. However, the two methods detailed above still have a good level of agreement with each other.

The γ-H2AX histone assay has been shown to be able to identify a recent partial body exposure of any dose, as well as any remaining foci several days after a high dose partial body exposure. Both the contaminated Poisson and Qdr methods have been used with this assay, although according to Horn, Barnard, and Rothkamm, the contaminated Poisson method was generally more accurate (judged by p values from Pearson's chi-square goodness of fit tests).

6.1.2 Full Count Distributions

Zero inflation is a phenomenon that is frequently related to overdispersion, and like overdispersion it can reduce the suitability of a Poisson fit. As discussed in Section 3.2.1, zero inflation occurs when the number of zeros in a sample significantly exceeds the expected amount. For count data, this amount is the expected number of zeros generated by a Poisson distribution with the same mean as the sample. Due to the relationship between them, distributions that account for overdispersion (e.g. compound Poisson and mixed Poisson distributions) will, to some extent, account for zero inflation too. This still generates an insufficient amount of zeros for many real life scenarios though, including agricultural, medical, and manufacturing ones. To account for this, we can use models that have been specifically designed to deal with zero-inflated count regression. Zero-inflated count models aim to account for the excess zeros by modelling data as a combination of two distributions: one that takes a single value at zero, then a count distribution for all non-zero values [27]. The two most commonly used regression models of this type are the zero-inflated Poisson (ZIP) model and the zero-inflated negative binomial (ZINB) model. Although not discussed earlier, these models can also be used for partial body irradiation scenarios.

To define a ZIP model, we take d blood samples from a healthy donor and irradiate them with several doses D_i, $i = 1, \ldots, d$. For each irradiated sample of n cells, let Y_{ij}, $i = 1, \ldots, d$, $j = 1, \ldots, n_i$ be our response variable which represents the observed number of dicentrics at dose level i for cell j. We can then define the ZIP regression model as:

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

where $0 \leq p_i \leq 1$ and $\lambda_i > 0$. The ZIP model's expectation and variance are defined by $E(Y_{ij}) = (1 - p_i)\lambda_i = \mu_i$ and $Var(Y_{ij}) = (1 - p_i)\lambda_i(1 + p_i\lambda_i)$, where λ_i refers to the mean of the underlying Poisson distribution and p_i is the mixture parameter (also referred to as the "zero-inflation parameter"). Both λ_i and p_i can depend on vectors of covariates [23]. As $Var(Y_{ij}) = \mu_i(1 + p_i\lambda_i) \geq \mu_i$, zero inflation can be considered to be a special form of overdispersion.

A ZIP model may provide a better fit for overdispersion that is linked to a large number of zeros with regards to the Poisson model. This type of zero-inflated regression model assumes that the observed zeros have two different origins: one group of zeros is produced at random by the Poisson distribution, whilst the other group are considered to be "structural". These so-called "structural" zeros, with proportion p_i, must be justified by the nature of the data, such as non-irradiated lymphocytes following a partial body exposure [23]. However, ZIP models may not be suitable for scenarios with multiple sources of excess zeros, such as densely ionising radiation. Here the ZINB regression model may be more appropriate. Introduced by Greene in 1994, the ZINB model is an extended version of the negative binomial model for excess zero count data. For situations where overdispersion is due to both an excess of zeros and heterogeneity of data, the ZINB regression model is often more suitable than the ZIP one.

The ZINB model with response variable Y_{ij} ($i = 1, \ldots, d$; $j = 1, \ldots, n_i$) has probability mass function:

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

where $\alpha > 0$ is an overdispersion parameter, and the index $c \in 0, 1$ gives the form of the underlying distribution. Oliveira et al. (2015) denote these two models as ZINB1 and ZINB2 respectively [23]. In general, the ZINB model has mean $E(Y_{ij}) = (1 - p_i)\lambda_i = \mu_i$ and variance $Var(Y_{ij}) = (1 - p_i)\lambda_i(1 + p_i\lambda_i + \alpha\lambda_i^c)$. The ZINB model reduces to the ZIP model as $\alpha \to 0$. This is analogous to the relationship between the negative binomial and Poisson distributions.

Zero-inflated models are not the only ones that have been considered to deal with overdispersed data. Aside from the Poisson and negative binomial distributions, Oliveira et al. (2015) consider four other types of distributions: the Neyman type A distribution, the univariate rth-order Hermite distributions, Poisson-inverse Gaussian distributions, and Pólya-Aeppli distributions.

6.2 Conclusions

In this thesis we have successfully analysed an unpublished internal factsheet from Public Health England [9] entitled "Dose and uncertainty estimation with the gamma-H2AX assay", which briefly and succinctly discusses suggested procedures to estimate doses from blood or lymphocyte samples using the γ-H2AX assay biomarker. We explored the techniques detailed on it with reference to the field of biodosimetry as a whole then applied them to two datasets from Public Health England. The first dataset, PHE (2017), our primary dataset, was used to demonstrate and illustrate the techniques and procedures described in the factsheet. The second dataset, PHE (2018), was then used to check that the methods described work for scenarios other than the first dataset. The methods described guide the user through calibration curve construction, dose estimation, and uncertainty estimation for a given dataset, including the introduction of a reference sampling ratio to validate the fitted calibration curves.

Overall the methods included on this factsheet not only work, but are of use when constructing and validating dose-response curves, as well as estimating exposed doses and the uncertainty on them. The linear dose-response curves fitted using both PHE datasets follow the general trend established in the literature (as shown in Section 5.3.2), namely that the value of the slope (α) coefficient decreases with increased time. The intercept (A) coefficient appears to remain fairly similar throughout for models fitted using the same data with little obvious consistent time-based effect. However, these values differ between datasets to an extent where we would not feel comfortable using them to assess the corresponding coefficients from other datasets. We only fitted quadratic models using the PHE (2017) dataset, but we did see a negative value quadratic (β) coefficient as expected, due to the saturation effect.

Using the summary statistics from the PHE (2017) fitted models, we were able to verify the relationship between Poisson and quasi-Poisson standard errors (and thus variances) in R. Quasi-Poisson standard errors can be calculated by multiplying the corresponding Poisson standard errors by the square root of the dispersion. Thus, for pa-

rameter estimates $\hat{\boldsymbol{\theta}}$, $SE_Q(\hat{\boldsymbol{\theta}}) = \sqrt{\phi} SE_P(\hat{\boldsymbol{\theta}})$ (Equation 3.9). Knowing this we can also calculate quasi-Poisson variances from Poisson ones with $\text{Var}_Q(\hat{\boldsymbol{\theta}}) = \phi \text{Var}_P(\hat{\boldsymbol{\theta}})$ (Equation 3.8). Our PHE (2017) fitted linear models had comparatively small standard errors associated with their coefficients, whilst the PHE (2018) standard errors were larger overall. This was expected though, and can be easily explained by our choice to fit a joint model to the data, as combining multiple fitted coefficients for the 4h and 24h slopes and intercepts also meant combining their standard errors. As well as this, there were a greater number of observed yields available for the 2017 data.

Calculating the variances associated with our fitted models' coefficients led us to successfully use the Delta Method to work out the uncertainty associated with a given dose. We realised that this equation could be rewritten as a quadratic with the yield y as the subject, but in the end this proved to be of little use. At first, we did consider the uncertainty on all eight of the Poisson and quasi-Poisson fits for the PHE (2017) dataset, although ultimately we decided to focus on the quasi-Poisson fits as they have larger variances (both for each coefficient and overall). We split the uncertainty equation up into "terms", the coefficients of each variable's variance σ^2 multiplied by the variance itself (see Section 4.4.3 for more details) to determine how much each coefficient contributed to the total uncertainty. From doing this with the quasi-Poisson fits at yields of 1 and 5 foci/cell (Figures D.2 and 4.1 respectively) we can clearly see that introducing a quadratic (or β) term vastly increases the uncertainty on the dose, potentially to levels where the uncertainty is so large as to be effectively useless. The total uncertainty is far lower for linear fits, and thus they are preferable to quadratic ones when fitting dose-response curves. This new plotting technique thus justifies the existing literature's preference for linear fits (and their lack of use of quadratic fits). The factsheet explicitly states that they do not consider the effects of the sampling time, but, at least with our PHE (2017) data, it is noticeable that the 24h fits (with their smaller slope coefficients) have greater uncertainty attached to them than the 1h ones, by roughly two orders of magnitude. This would need checking with other datasets before any firm conclusions are drawn, though.

We were ultimately able to use reference samples to validate some fitted curves, calculate reference sampling ratios, and use those ratios to adjust our previously fitted models. Upon calculating yields for our existing models at the two control dose points, 0Gy and 1.5Gy, we discovered that only the two 1.5Gy yields for the 24h fits (linear and quadratic) were within the required 30% discrepancy interval for curve validation. However, as the 1.5Gy is a more important control dose than 0Gy, we could still technically validate the two 24h curves. We were then successfully able to calculate the reference sampling ratio

and adjust the dose equations accordingly.

However, understandably, the work done for this thesis was not without some issues, especially when it came to introducing the second dataset (PHE (2018)). This dataset, unlike our original one, contained data that was scored by two different scientists, and we performed a two sample t test that showed a significant difference between their scored counts. This presented an issue as we then had to normalise and correct the data. As well as this, we had to ensure that the data points we chose to use as reference samples were scored by the same person to avoid scorer issues affecting the calculations. Our individual corrected data points cannot be directly compared to the integer counts from the other dataset as they are not actually scored cells. The correction part of this process (detailed in Section 5.2.2) was followed by aggregating the data to give 22 data points to fit models with. We faced some issues with fitting the models, as the twelve 0Gy data points had the same corrected value, but did end up successfully fitting joint models for this data which did prove to be useful. If we were to try and fit this data using GLMs again it may be worth trying to use the full corrected data without aggregation, to see if it would be possible to do so, then compare any dose-response curves fitted to those fitted using the aggregated data.

It is notable that our PHE (2017) linear calibration curves could not be fully validated using the method as written in the factsheet due to none of the 0Gy yields being within the 30% discrepancy interval. The factsheet requires a discrepancy of less than 30% for both control points in order for a curve to be validated. However, we were able to validate two of the curves (24h linear and quadratic) on a technicality, by arguing that the 1.5Gy control point is of greater importance. (We did also construct alternative curves instead if needed.) Some of our issues with the reference samples may have been due to using readings from the PHE (2018) dataset as control samples. These may not have been scored by the same person and in identical conditions, and there were noted issues with multiple scorers in this dataset.

Although we were able to draw interesting and useful conclusions from our Delta Method/uncertainty graphs for the PHE (2017) dataset, it should be noted that this particular type of graph is a new technique that requires testing further to see if it is of use in general for this method, rather than just for this dataset. Theoretically it is applicable to other datasets with dose-response calibration curves, but there is a significant difference between theory and practice.

We also had a slight issue with the formulas listed on the factsheet. PHE suggested estimating the standard error on the yield, σ_y, with the equation $\sigma_y = \frac{\sqrt{y}}{n}$ in the Poisson

case.

From a practical perspective, there are some obvious issues with the sample timepoints used in this thesis. A 1h timepoint (as seen in the PHE (2017) dataset) may be useful statistically, but in a real life situation it is very unlikely that a blood sample would be successfully taken an hour after a suspected radiation exposure. The factsheet suggests using 4h and 24h as default timepoints, but flags up another potential issue: that the calibration curves and estimates calculated will only work well for samples taken near these timepoints. If this is not the case, further calibration curves should be constructed if possible, or a different assay should be used.

Radiation biodosimetry as a research area is primarily concerned with working out how much radiation a person has been exposed to as accurately as possible. Methods that can process a large number of samples relatively quickly are preferable for potential triage situations, although many real life radiation exposures affect smaller numbers of people. Within that context, our aim was to see if proposed standard procedure by PHE, a UK Government agency of the Department of Health and Social Care, would be feasible for dose estimation and thus how useful the factsheet detailing it is. Overall we discovered that yes, the methods did work, and the factsheet is of use to people wishing to analyse γ-H2AX assay data. The standard error equation stated only works for Poisson fits, and applying the 30% discrepancy interval validation benchmark to intercept coefficients as well as slope and quadratic ones can be problematic. Despite these (relatively minor) issues, the factsheet is a good framework overall for analysing foci data. Analysing this factsheet and testing its methods also led us to learn that linear fits are preferable to quadratic ones for dose-response curves, and that quasi-Poisson models are of more use than Poisson ones for fitting dose-response models. This work also helped construct a unified statistical methodology for calibration curve estimation, dose estimation, and uncertainty quantification with the γ-H2AX assay, and this fed into a published paper [10].

If we were to potentially extend this project, applying the techniques to more datasets for verifications would be useful, especially to verify the effectiveness of the new uncertainty graphs that we created. We could also look more at various validation criteria for dose-response curves using reference samples, including applying different ones to each coefficient, to see if 30% is truly the best option possible.

List of Figures

List of Tables

Appendix A

Dataset Summaries

A.1 First Dataset - PHE (2017)

	1h						24h						
	0 Gy	0.05	0.1	0.25	0.5	1	0	0.05	0.1	0.25	0.5	1	4
H02	0.08				7.19								7.65
H02	0.06					13.58							
H02						13.52						2.08	
H02							0.14					2.22	
H02 (M2)	0.00				8.40	11.90	0.10				1.60	3.70	7.70
H02	0.14				6.86								
H02					7.52								
H03 (F3)	0.02				6.92	11.46	0.22				2.14	2.99	8.50
H07	0.05				7.35								
H07	0.26												
H07	0.04	0.88	1.52	3.58	6.62	10.62	0.08	0.32	0.30	0.46	1.02	1.78	
H07	0.10	0.76	2.02	4.26	7.02	13.46	0.08	0.22	0.38	0.72	1.16	1.78	
H08	0.12					12.38							
H08						12.26						2.36	
H08							0.18					2.54	
H09	0.08				7.30	9.94							
H09						9.86						2.38	
H09							0.04					2.42	
H09	0.18												
H09	0.04	0.70	1.44	3.26	7.02	9.32	0.04	0.50	0.44	1.66	1.78	1.74	
H09	0.08	1.20	2.08	4.02	6.70	12.12	0.06	0.28	0.42	1.36	1.12	1.78	
H12	0.11												7.69
H12	0.54				7.60								7.66
H12	0.04												
H12	0.08	0.80	1.30	4.12	6.80	9.76	0.08	0.30	0.52	0.60	1.84	1.56	
H12	0.06	0.96	1.60	3.22	6.62	11.36	0.12	0.26	0.34	0.78	0.94	1.72	
H14					7.48								
H15	0.34				6.38								8.52
H15					6.04								7.98
H15	0.54												

Figure A.1: A screenshot of part of our first dataset, showing the raw data.

A.2 Second Dataset - PHE (2018)

Slide: 1h 0%-100% 0,75Gy – Well 1: 0%					
total	Cell	Foci		Cells in total	200
1	1	1		Foci in total	143
1	2	0		Avrage Foci per Cell	0.715
1	3	0			
1	4	0			
1	5	2			
1	6	0			
1	7	0			
1	8	0			
1	9	0			
1	10	2			
1	11	0			
1	12	0			
1	13	0			
1	14	7			
1	15	0			
1	16	0			
1	17	0			
1	18	0			
1	19	0			
1	20	1			
1	21	0			
1	22	0			
1	23	0			
1	24	0			
1	25	0			
1	26	0			
1	27	0			

Figure A.2: A screenshot of part of our second dataset, showing the raw data being inputted. The scenario shown in part here is actually 0Gy data, due to the 0% stated exposure.

	Dose	Timepoint	Percentage	Scorer	Result	Result (normalised so 0 is avg. 0)
0.75	0.75	1	0	1	1	0.285
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	2	1.285
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	2	1.285
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	7	6.285
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	1	0.285
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715
	0.75	1	0	1	0	-0.715

Figure A.3: A screenshot of part of our second dataset. The yellow column shows the percentage of the sample that was exposed to the stated dose. In this case this is 0, so we actually have 0Gy data here. The "Result" column gives the scored foci count per cell. The right hand column shows initial normalisation by PHE scientists.

Appendix B

List of Partial Derivatives

B.1 Dose Equations

1. $D = \frac{(y-A)}{\alpha}$ (linear, not including reference sampling ratio)
2. $D = \frac{-\alpha+\sqrt{\alpha^2+4\beta(y-A)}}{2\beta}$ (quadratic, not including reference sampling ratio)
3. $D = \frac{(y-A)}{\alpha}r$ (linear, including reference sampling ratio)
4. $D = \frac{-\alpha+\sqrt{\alpha^2+4\beta(y-A)}}{2\beta}r$ (quadratic, including reference sampling ratio)

B.2 Not Including Reference Sampling Ratio

B.2.1 Linear

Dose equation:

$$D_L = \frac{(y-A)}{\alpha}$$

Partial derivatives:

$$\frac{\partial D_L}{\partial A} = \frac{-1}{\alpha}$$

$$\frac{\partial D_L}{\partial y} = \frac{1}{\alpha}$$

$$\frac{\partial D_L}{\partial \alpha} = \frac{A-y}{\alpha^2}$$

B.2.2 Quadratic

Dose equation:

$$D_Q = \frac{-\alpha + \sqrt{\alpha^2 + 4\beta\,(y - A)}}{2\beta}$$

Partial derivatives:

$$\frac{\partial D_Q}{\partial A} = \frac{1}{\sqrt{\alpha^2 + 4\beta(y - A)}}$$

$$\frac{\partial D_Q}{\partial y} = \frac{-1}{\sqrt{\alpha^2 + 4\beta(y - A)}}$$

$$\frac{\partial D_Q}{\partial \alpha} = \frac{1}{2\beta}\left(\frac{\alpha}{\sqrt{\alpha^2 + 4\beta(y - A)}} - 1\right)$$

$$\frac{\partial D_Q}{\partial \beta} = \frac{1}{\beta}\left(\frac{y - A}{\sqrt{\alpha^2 + 4\beta(y - A)}} + \frac{\alpha}{2\beta} - \frac{\sqrt{\alpha^2 + 4\beta(y - A)}}{2\beta}\right)$$

B.3 Including Reference Sampling Ratio

B.3.1 Linear

Dose equation:

$$D_{Lr} = \frac{(y - A)}{\alpha} r$$

Partial derivatives:

$$\frac{\partial D_{Lr}}{\partial A} = \frac{-r}{\alpha}$$

$$\frac{\partial D_{Lr}}{\partial y} = \frac{r}{\alpha}$$

$$\frac{\partial D_{Lr}}{\partial \alpha} = \frac{r(A - y)}{\alpha^2}$$

$$\frac{\partial D_{Lr}}{\partial r} = \frac{y - A}{\alpha} = D_L$$

B.3.2 Quadratic

Dose equation:

$$D_{Qr} = \frac{-\alpha + \sqrt{\alpha^2 + 4\beta\,(y - A)}}{2\beta} r$$

Partial derivatives:

$$\frac{\partial D_{Qr}}{\partial A} = \frac{r}{\sqrt{\alpha^2 + 4\beta(y - A)}}$$

$$\frac{\partial D_{Qr}}{\partial y} = \frac{-r}{\sqrt{\alpha^2 + 4\beta(y - A)}}$$

$$\frac{\partial D_{Qr}}{\partial \alpha} = \frac{r}{2\beta}\left(\frac{\alpha}{\sqrt{\alpha^2 + 4\beta(y - A)}} - 1\right)$$

$$\frac{\partial D_{Qr}}{\partial \beta} = \frac{r}{\beta}\left(\frac{y - A}{\sqrt{\alpha^2 + 4\beta(y - A)}} + \frac{\alpha}{2\beta} - \frac{\sqrt{\alpha^2 + 4\beta(y - A)}}{2\beta}\right)$$

$$\frac{\partial D_{Qr}}{\partial r} = \frac{-\alpha + \sqrt{\alpha^2 + 4\beta(y - A)}}{2\beta} = D_Q$$

Appendix C

Code

C.1 General Case

```
#linear y reformulation
s2D.liy<- function(y,fit,n,count){
 A1<- fit$coefficients[[1]]
 a1<- fit$coefficients[[2]]
 d1li<-summary(fit)$dispersion
 s2A1<- vcov(fit)[1,1]
 s2a1<- vcov(fit)[2,2]
 s2y1<- d1li*y/n
 k1<- s2a1/a1^4
 k2<- (-2*A1*s2a1)/a1^4
 k3<- a1^-2*(s2A1+s2y1+A1^2*s2a1/a1^2)
 k=c(k1,-k2,k3)
 names(k)=c("y^2","y","constant")
 barplot(k, xlab="Coefficients, listed by order of y",ylab="Value",
 main="Linear,y")
 sigma2D1q<- k1*y^2+k2*y+k3
 print(k)
 return(sigma2D1q)
}

#linear
s2D.li<- function(y,fit,n,count){
```

```
 A1<- fit$coefficients[[1]]
 a1<- fit$coefficients[[2]]
 d1li<-summary(fit)$dispersion
 s2A1<- vcov(fit)[1,1]
 s2a1<- vcov(fit)[2,2]
 s2y1<- d1li*y/n
 q1<-s2A1/a1^2
 q2<-s2y1/a1^2
 q3<-(A1-y)^2*s2a1*a1^-4
 q=c(q1,q2,q3)
 names(q)=c("A","y","alpha")
 barplot(q, xlab=expression(paste("Coefficients of ", sigma^2)),ylab="Value",
 main="Linear")
 sigma2D1<- q1+q2+q3
 print(q)
 return(sigma2D1)
}

#quadratic
s2D.qu<- function(y,fit,n,count){
 A3<- fit$coefficients[[1]]
 a3<- fit$coefficients[[2]]
 b3<- fit$coefficients[[3]]
 d1qu<-summary(fit)$dispersion
 s2A3<- vcov(fit)[1,1]
 s2a3<- vcov(fit)[2,2]
 s2b3<- vcov(fit)[3,3]
 s2y3<- d1qu*y/n
 z3<- a3^2+4*b3*(y-A3)
 m1<- s2A3/z3
 m2<- s2y3/z3
 m3<- (-1+a3/z3^0.5)^2*s2a3*0.25*b3^-2
 m4<- ((y-A3)^2/z3+a3*(y-A3)*b3^-1*z3^-0.5+(y-A3)/b3+a3*z3^0.5*b3^-2*0.5+
 a3^2*b3^-2*0.25+z3*b3^-2*0.25)*s2b3*b3^-2
 m=c(m1,m2,m3,m4)
```

```
 names(m)=c("A","y","alpha","beta")
 barplot(m, xlab=expression(paste("Coefficients of ", sigma^2)),ylab="Value",
 main="Quadratic")
 sigma2D3q<- m1+m2+m3+m4
 print(m)
 return(sigma2D3q)
}
```

C.2 Supplementary Code For Section 4.5.2

```
#values for a yield of 5 foci/cell (with code)
n=500
Y=5
#linear, 1h
s2D1qdx<- function(y){
 A1<- fit1q.li$coefficients[[1]]
 a1<- fit1q.li$coefficients[[2]]
 d1li<-summary(fit1q.li)$dispersion
 s2A1<- c1q.li$cov[1,1]
 s2a1<- c1q.li$cov[2,2]
 s2y1<- d1li*y/n
 q1<-s2A1/a1^2
 q2<-s2y1/a1^2
 q3<-(A1-y)^2*s2a1*a1^-4
 q=c(q1,q2,q3)
 names(q)=c("A","y","alpha")
 barplot(q, xlab=expression(paste("Coefficients of ", sigma^2)),ylab="Value",
 main="Quasi-poisson, linear, 1h")
 sigma2D1<- q1+q2+q3
 print(q)
 return(sigma2D1)
}

#linear, 24h
```

```
s2D2qdx<- function(y){
 A2<- fit24q.li$coefficients[[1]]
 a2<- fit24q.li$coefficients[[2]]
 d24li<-summary(fit24q.li)$dispersion
 s2A2<- c24q.li$cov[1,1]
 s2a2<- c24q.li$cov[2,2]
 s2y2<- d24li*y/n
 q1<-s2A2/a2^2
 q2<-s2y2/a2^2
 q3<-(A2-y)^2*s2a2*a2^-4
 q=c(q1,q2,q3)
 names(q)=c("A","y","alpha")
 barplot(q, xlab=expression(paste("Coefficients of ", sigma^2)),ylab="Value",
 main="Quasi-poisson, linear, 24h")
 sigma2D2<- q1+q2+q3
 print(q)
 return(sigma2D2)
}

#quadratic, 1h
s2D3qd<- function(y){
 A3<- fit1q.qu$coefficients[[1]]
 a3<- fit1q.qu$coefficients[[2]]
 b3<- fit1q.qu$coefficients[[3]]
 d1qu<-summary(fit1q.qu)$dispersion
 s2A3<- c1q.qu$cov[1,1]
 s2a3<- c1q.qu$cov[2,2]
 s2b3<- c1q.qu$cov[3,3]
 s2y3<- d1qu*y/n
 z3<- a3^2+4*b3*(y-A3)
 m1<- s2A3/z3
 m2<- s2y3/z3
 m3<- (-1+a3/z3^0.5)^2*s2a3*0.25*b3^-2
 m4<- ((y-A3)^2/z3+a3*(y-A3)*b3^-1*z3^-0.5+(y-A3)/b3+a3*z3^0.5*b3^-2*0.5+
 a3^2*b3^-2*0.25+z3*b3^-2*0.25)*s2b3*b3^-2
```

```
 m=c(m1,m2,m3,m4)
 names(m)=c("A","y","alpha","beta")
 barplot(m, xlab=expression(paste("Coefficients of ", sigma^2)),ylab="Value",
 main="Quasi-poisson, quadratic, 1h")
 sigma2D3q<- m1+m2+m3+m4
 print(m)
 return(sigma2D3q)
}

#quadratic, 24h
s2D4qd<- function(y){
 A4<- fit24q.qu$coefficients[[1]]
 a4<- fit24q.qu$coefficients[[2]]
 b4<- fit24q.qu$coefficients[[3]]
 d24qu<-summary(fit24q.qu)$dispersion
 s2A4<- c24q.qu$cov[1,1]
 s2a4<- c24q.qu$cov[2,2]
 s2b4<- c24q.qu$cov[3,3]
 s2y4<- d24qu*y/n
 z4<- a4^2+4*b4*(y-A4)
 m1<- s2A4/z4
 m2<- s2y4/z4
 m3<- (-1+a4/z4^0.5)^2*s2a4*0.25*b4^-2
 m4<- ((y-A4)^2/z4+a4*(y-A4)*b4^-1*z4^-0.5+(y-A4)/b4+a4*z4^0.5*b4^-2*0.5+
 a4^2*b4^-2*0.25+z4*b4^-2*0.25)*s2b4*b4^-2
 m=c(m1,m2,m3,m4)
 names(m)=c("A","y","alpha","beta")
 barplot(m, xlab=expression(paste("Coefficients of ", sigma^2)),ylab="Value",
 main="Quasi-poisson, quadratic, 24h")
 sigma2D4q<- m1+m2+m3+m4
 print(m)
 return(sigma2D4q)
}
par(mfrow=c(2,2))
```

Appendix D

Plots

All dose estimates in this appendix are given to four decimal places.

D.1 Yield of 5 foci/cell, Continued From Section 4.5.2

Fit type	Time	Dose Estimate (Gy)
Linear	1h	0.3877
	24h	2.4884
Quadratic	1h	0.3477
	24h	2.3675

Table D.1: Dose estimates for a yield of 5 foci/cell in Gray, to four decimal places.

D.2 Yield of 1 foci/cell

Fit type	Time	Dose Estimate (Gy)
Linear	1h	0.0692
	24h	0.4236
Quadratic	1h	0.0581
	24h	0.3834

Table D.2: Dose estimates for a yield of 1 foci/cell in Gray, to four decimal places.

1h:

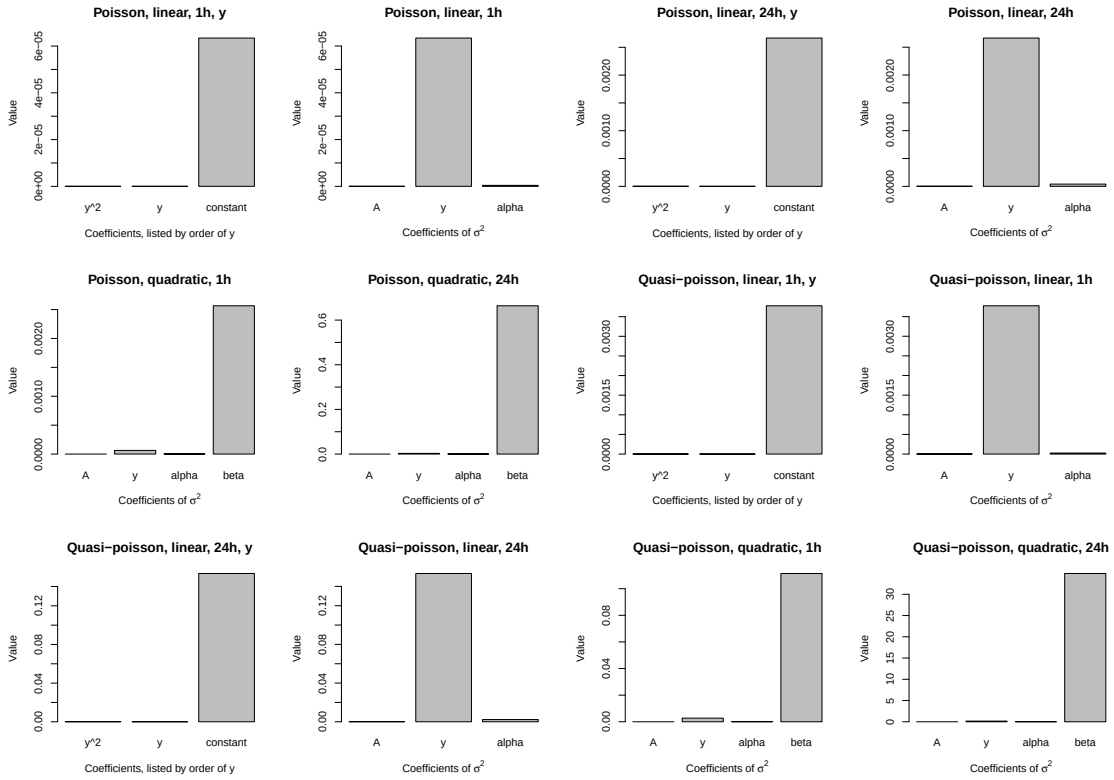

Figure D.1: A series of twelve plots showing the contribution to the overall uncertainty for each "term" in a given scenario with a yield of 1 foci/cell.

Linear:

$$D = \frac{1 - 0.130838}{12.558857} = 0.0692Gy$$

Quadratic:

$$D = \frac{-15.507491 + \sqrt{15.507491^2 + 4 \times -4.169313\,(1 - 0.112397)}}{2 \times -4.169313} = \frac{-15.507491 + \sqrt{225.6795}}{-8.338626}$$

$$D = 0.0581Gy$$

24h:

Linear:

$$D = \frac{1 - 0.179394}{1.937251} = 0.4236Gy$$

Quadratic:

$$D = \frac{-2.275551 + \sqrt{2.275551^2 + 4 \times -0.094352\,(1 - 0.141415)}}{2 \times -0.094352} = \frac{-2.275551 + \sqrt{4.854096}}{-0.188704}$$

$$D = 0.3834Gy$$

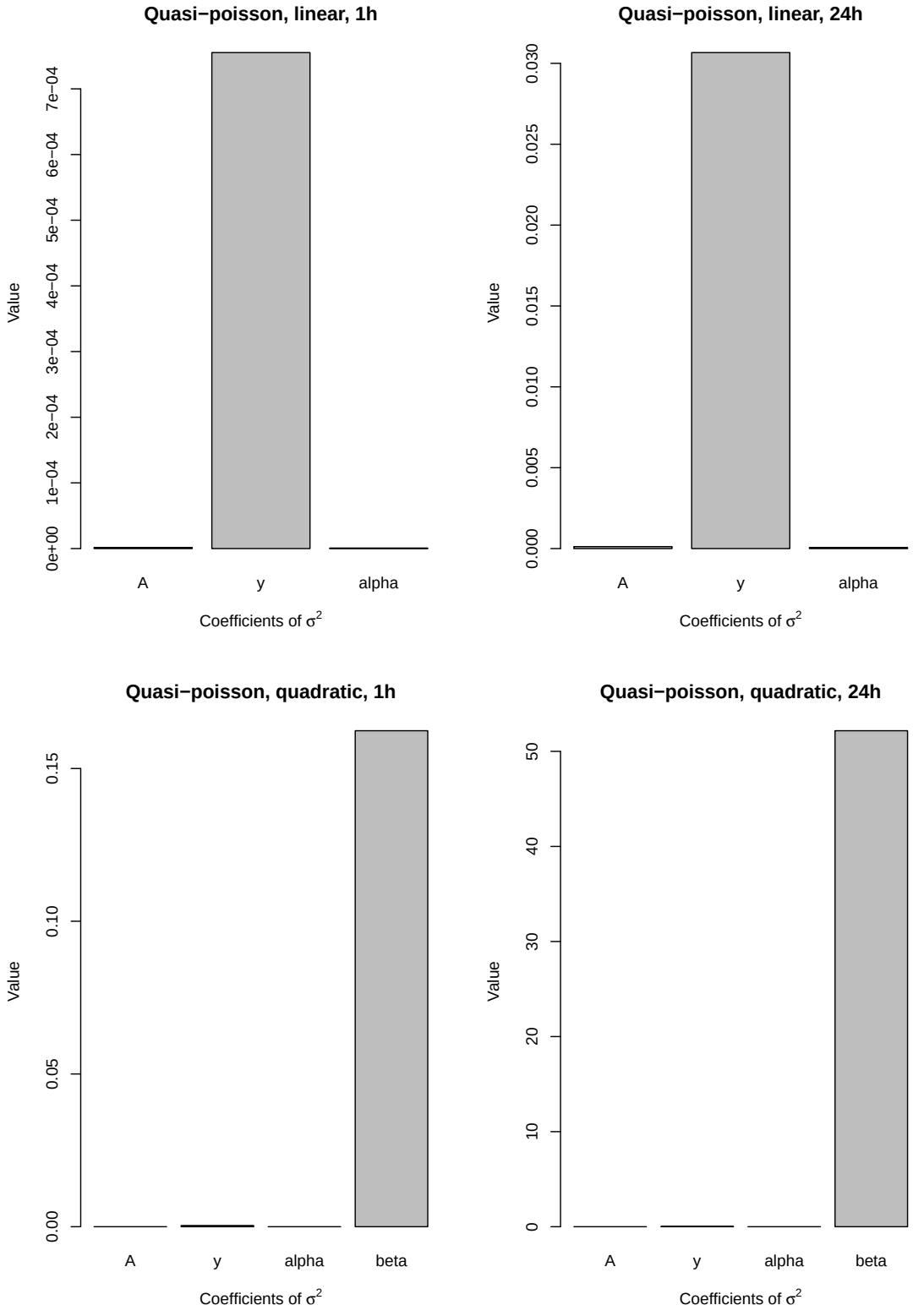

Figure D.2: A series of four plots showing the contribution to the overall uncertainty for each "term" in a given scenario, accompanying the code given in Appendix C for a yield of 1 foci/cell. This figure is the 1 foci/cell equivalent of Figure 4.1

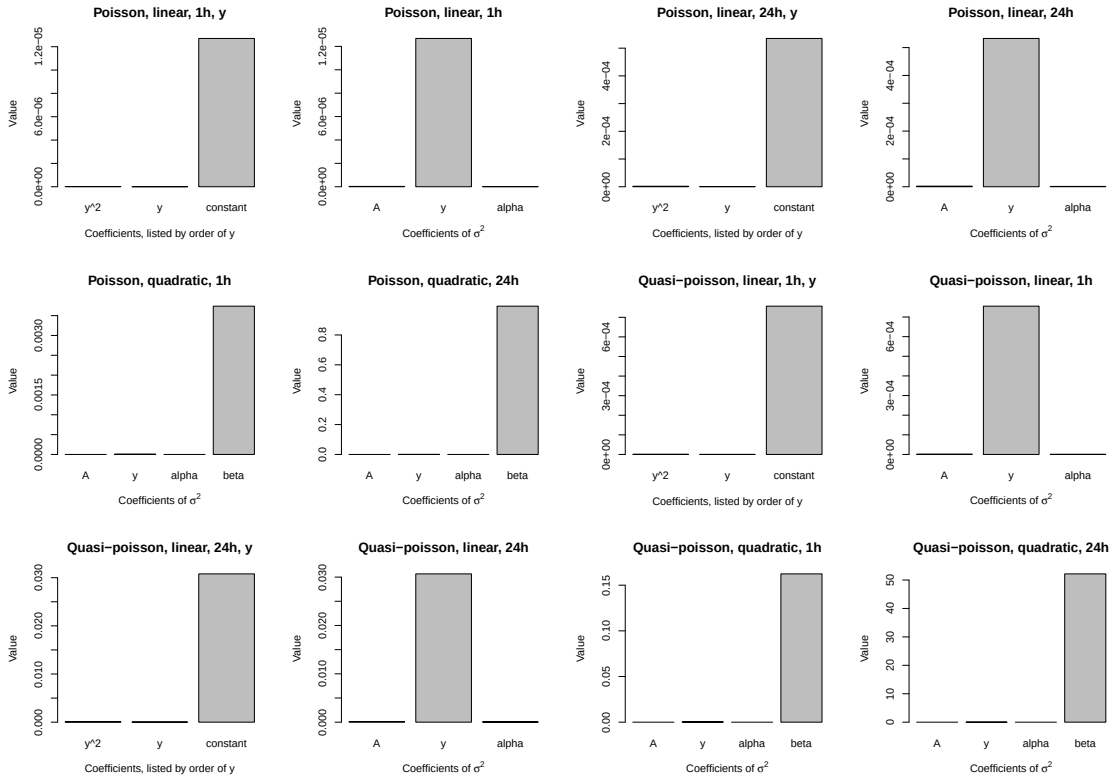

Figure D.3: A series of twelve plots showing the contribution to the overall uncertainty for each "term" in a given scenario with a yield of 1 foci/cell.

Bibliography

[1] Jochen Einbeck. *Radiation dosimetry through statistical analysis of biomarkers*. Presentation for Undergraduate Colloquium. Durham: Durham University.

[2] M Sun et al. *Doses in Radiation Accidents Investigated by Chromosomal Aberration Analysis XXV. Review of cases investigated, 2006-2015*. Public Health England, 2016.

[3] Jayne Moquet et al. "The second gamma-H2AX assay inter-comparison exercise carried out in the framework of the European biodosimetry network (RENEB)". In: *International Journal of Radiation Biology* 93.1 (2017), pp. 58–64.

[4] M Szłuińska, A A Edwards, and D C Lloyd. *Statistical Methods for Biological Dosimetry*. type. Health Protection Agency, Centre for Radiation, Chemical and Environmental Hazards, Radiation Protection Division, 2005.

[5] International Atomic Energy Agency. *Cytogenetic Dosimetry: Applications in Preparedness for and Response to Radiation Emergencies*. Austria: IAEA, 2011, pp. 1–100.

[6] Jochen Einbeck et al. "On the use of random effect models for radiation biodosimetry". In: *Extended Abstracts Fall 2015*. Vol. 8. Trends in Mathematics. Basel: Birkhäuser, Cham, pp. 89–94.

[7] Kai Rothkamm et al. "Manual versus automated γ-H2AX foci analysis across five European laboratories: Can this assay be used for rapid biodosimetry in a large scale radiation accident?" In: *Mutation Research* 756 (2013), pp. 170–173.

[8] *Bulletin*. Newsletter. Realising the European Network of Biodosimetry, 2015.

[9] Elizabeth Ainsbury. *Dose and uncertainty estimation with the gamma-H2AX assay*. Unpublished Factsheet. Public Health England, 2016.

[10] Jochen Einbeck et al. "A statistical framework for radiation dose estimation with uncertainty quantification from the γ-H2AX assay". In: *PLoS One* 13.11 (2018).

[11] Aida Muslimovic, Pegah Johansson, and Ola Hammarsten. "Measurement of H2AX Phosphorylation as a Marker of Ionizing Radiation Induced Cell Damage". In: *Current Topics in Ionizing Radiation Research*. 2012. ISBN: 978-953-51-0196-3.

[12] Dane Avondoglio et al. "High throughput evaluation of gamma-H2AX". In: *Radiation Oncology* 4.31 (2009).

[13] Emmy P. Rogakou et al. "DNA Double-stranded Breaks Induce Histone H2AX Phosphorylation on Serine 139". In: *The Journal of Biological Chemistry* 273.10 (1998), pp. 5858–5868.

[14] Jochen Einbeck. *The H2AX-histone as a radiation biomarker - modelling and dose estimation*. Presentation for Mini Workshop on Protein Structure Prediction and Modelling. Durham University.

[15] Kai Rothkamm and Simon Horn. "γ-H2AX as protein biomarker for radiation exposure". In: *Annali dell'Istituto Superiore Di Sanità* 45.3 (9), pp. 265–271.

[16] Jayne Moquet, Stephen Barnard, and Kai Rothkamm. "Gamma-H2AX biodosimetry for use in large scale radiation incidents: comparison of a rapid '96 well lyse/fix' protocol with a routine method". In: *PeerJ* (2014). Ed. by Sren Bentzen.

[17] Jochen Einbeck. *Radiation dose estimation through the gamma-H2AX protein*. Presentation for BSI Research Showcase. Durham: Durham University.

[18] K. Rothkamm et al. "Laboratory Intercomparison on the γ-H2AX Foci Assay". In: *Radiation Research* 180.2 (2013), pp. 149–155.

[19] Norman L. Johnson, Adrienne W. Kemp, and Samuel Kotz. *Univariate Discrete Distributions*. 3rd. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley and Sons, Inc, 2005. Chap. 4. ISBN: 978-0-471-27246-5.

[20] A. Colin Cameron and Pravin K. Trivedi. *Regression Analysis of Count Data*. Econometric Society Monographs. Cambridge: Cambridge University Press, 1998. Chap. 3, pp. 59–85. ISBN: 0-521-63567-5.

[21] A. Hilali et al. "An Appraisal of the Value of the Contaminated Poisson Method to Estimate the Dose Inhomogeneity in Simulated Partial-Body Exposure". In: *Radiation Research* 128.1 (1991), pp. 108–111.

[22] Simon Horn, Stephen Barnard, and Kai Rothkamm. "Gamma-H2AX-Based Dose Estimation for Whole and Partial Body Radiation Exposure". In: *PLoS One* 6.9 (2011).

[23] Mara Oliveira et al. “Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study”. In: *Biometrical Journal* 58.2 (2015), pp. 259–279.

[24] Kai Rothkamm et al. “DNA Damage Foci: Meaning and Significance”. In: *Environmental and Molecular Mutagenesis* 56.6 (2015), pp. 491–504.

[25] Jay M. Ver Hoef and Peter L. Boveng. “Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?” In: *Ecology* 88.11 (2007), 27662772.

[26] Ludwig Fahrmeir and Gerhard Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. 2nd. Springer Series in Statistics. Springer-Verlag, 2001, pp. 55–60. ISBN: 0-387-95187-3.

[27] Adam Errington. “Investigating the Role of Zero-Inflated Models in Relation to Dose Estimation”. MSc. University of Durham, 2017.

[28] Elizabeth Ainsbury et al. “Uncertainty of fast biological radiation dose assessment for emergency response scenarios”. In: *International Journal of Radiation Biology* 93.1 (2017), pp. 127–135.

[29] Yuqi Gao. “γ-H2AX-based Dose Estimation via Standard Methodology in Dicentric Assay”. Master of Science Dissertation. University of Durham, 2017.

[30] Joint Committee For Guides In Metrology. *Evaluation of measurement data Guide to the expression of uncertainty in measurement*. 1st. 100.