Adam Errington*, Jochen Einbeck, Jonathan Cumming, Ute Rössler and David Endesfelder

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

https://doi.org/10.1515/ijb-2020-0079
Received May 29, 2020; accepted April 21, 2021; published online May 7, 2021

Abstract: For the modelling of count data, aggregation of the raw data over certain subgroups or predictor configurations is common practice. This is, for instance, the case for count data biomarkers of radiation exposure. Under the Poisson law, count data can be aggregated without loss of information on the Poisson parameter, which remains true if the Poisson assumption is relaxed towards quasi-Poisson. However, in biodosimetry in particular, but also beyond, the question of how the dispersion estimates for quasi-Poisson models behave under data aggregation have received little attention. Indeed, for real data sets featuring unexplained heterogeneities, dispersion estimates can increase strongly after aggregation, an effect which we will demonstrate and quantify explicitly for some scenarios. The increase in dispersion estimates implies an inflation of the parameter standard errors, which, however, by comparison with random effect models, can be shown to serve a corrective purpose. The phenomena are illustrated by γ-H2AX foci data as used for instance in radiation biodosimetry for the calibration of dose-response curves.

Keywords: heterogeneity; overdispersion; quasi-Poisson; radiation biomarker; random effect.

1 Introduction

The aggregation of count data prior to analysis or modelling is a very common procedure in several fields, including for instance the aggregation of clickstream data in e-commerce [1], or of species counts in ecology [2]. Furthermore, in biodosimetry, which is the context in which this paper is set, it is common to aggregate counts of certain biomarkers, such as chromosomal aberrations, over samples of blood cells, and use the aggregated count for the estimation of dose-response curves, or the estimation of dose given an existing curve.

A particular protein-based biomarker, based on the γ-H2AX histone, has motivated this work. H2AX is a key factor in the DNA damage and repair mechanisms. It is recruited to damage sites, which in turn recruit other DNA repair machinery. DNA is normally wrapped around a core histone molecule forming the nucleosome complex. Histone cores are made up of individual histone proteins: H2A, H2B, H3 and H4. The H2A protein family has the greatest number of variants including H2A1, H2A2, H2AX and H2AZ. Depending on the cell type, H2AX constitutes 2–20% of the H2A protein. The H2AX histone is a DNA-repair protein; that is, once a cell gets exposed to ionising radiation and a double-strand break (DSB) has occurred, it coordinates the repair of the damaged DNA and in this process phosphorylates, becoming γ-H2AX [3]. This phosphorylation leads, after addition of fluorophore-labelled antibodies, to fluorescent dots which can be counted under a microscope.

***Corresponding author**: **Adam Errington**, Department of Mathematical Sciences, Durham University, Durham, UK, E-mail: adamerrington22@hotmail.com. https://orcid.org/0000-0003-3728-349X
Jochen Einbeck and Jonathan Cumming, Department of Mathematical Sciences, Durham University, Durham, UK
Ute Rössler and David Endesfelder, Bundesamt für Strahlenschutz (BfS), Oberschleissheim, Germany

The suitability of this histone as a biomarker for DSBs [4, 5], and by extension, ionising radiation exposure [6–9], has long been established in the literature. However, statistical work to quantify this relationship and facilitate the actual dose estimation has only been carried out quite recently [10–12]. It should be noted that γ-H2AX foci data is not only used for biological dosimetry but much more prominently for several research questions in radiation biology [13–16].

In order to establish dose-response calibration curves, laboratory experiments are carried out where blood samples are exposed to known degrees of radiation. The data arising from a series of such experiments conducted at the Bundesamt für Strahlenschutz (BfS), Germany, are displayed in Figures 1 and 2. For the production of the data, whole blood samples were irradiated with one of six design doses (0.1, 0.2, 0.3, 0.4, 0.5 and 1 Gy), always with 195 kV X-radiation. One hour after exposure, blood samples consisting of approximately 2000 cells were then placed on slides under an immunoflourescence microscope, and the number of foci on each slide was counted in a semi-automatic way using MetaCyte software. (Additional information on the generation of these data set is deferred to Appendix A).

In total, measurements from 116 slides are available, corresponding to a total of 233,220 frequencies of foci per cell. Figure 1 gives an excerpt of the raw data, in the form of a frequency distribution of foci counts for three specific slides. One sees clearly how the distribution of the foci counts is shifted to the right for increasing doses, underlining their suitability as a radiation biomarker (note again that all cells on a given slide always share the same design dose). The full data set is displayed in Figure 2 in aggregated form, with each point corresponding to the mean foci count for a specific slide. From this one can deduce some sort of empirical dose-response relationship, which appears roughly linear over a considerable dose range, noting a saturation effect [17] for higher doses.

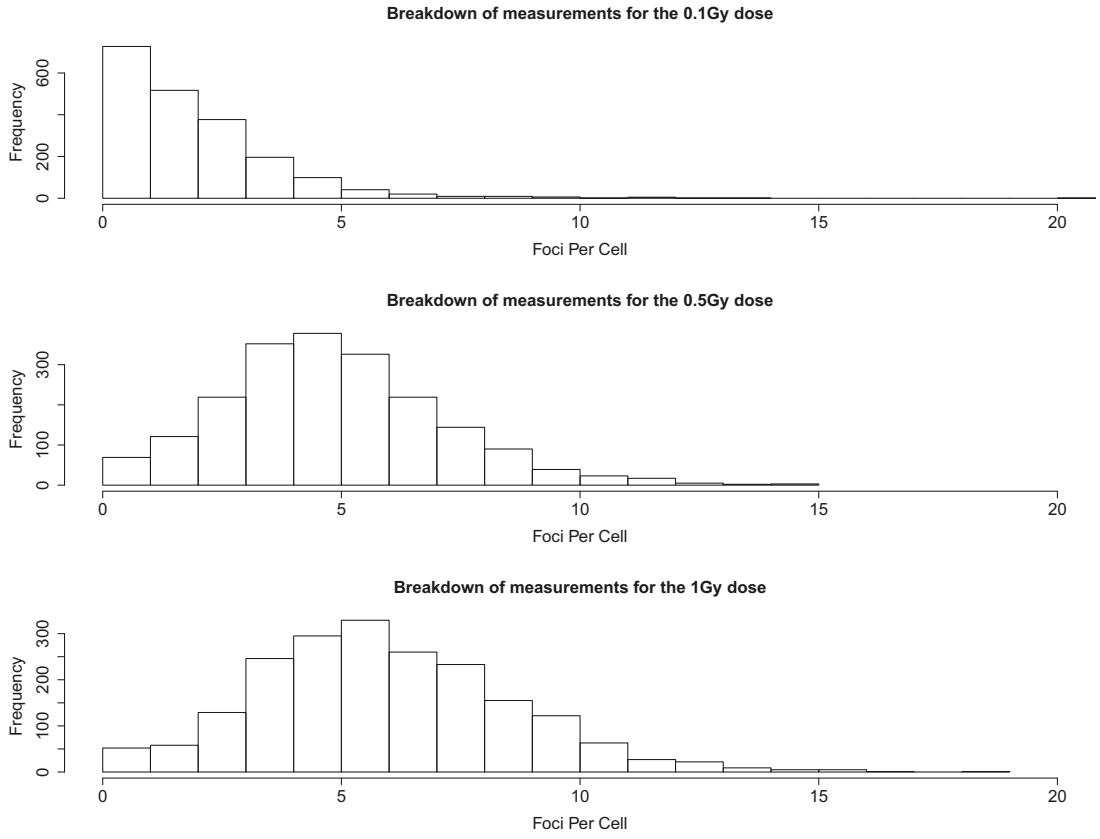

Figure 1: Distribution of the number of observed foci, for three selected slides with dose levels 0.1, 0.5 and 1 Gy, respectively. As one reaches a higher level of dose, the number of foci tends to increase, yielding a reduced percentage of zero counts.

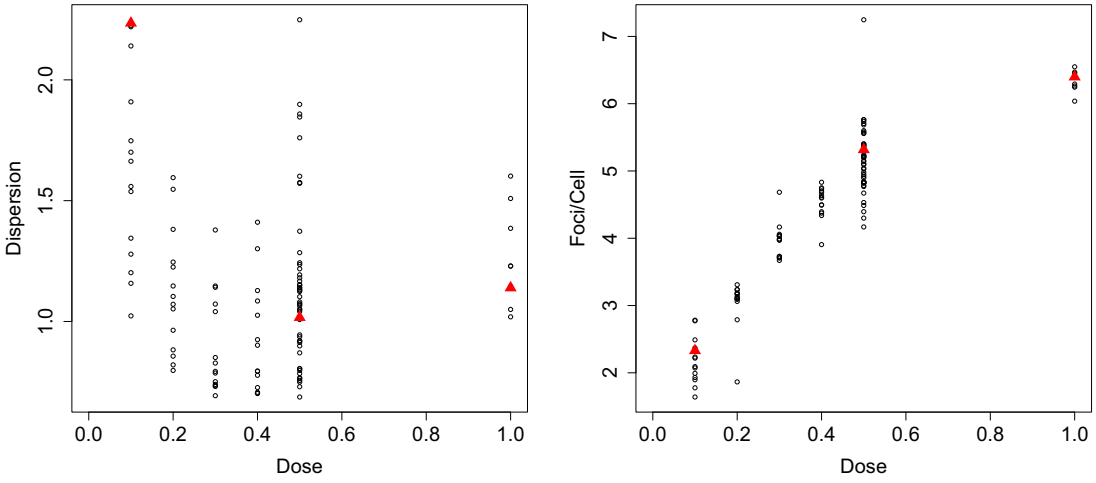

Figure 2: Slide-wise dispersions (left) and foci yields (right) recorded for various levels of dose. The three points highlighted as triangles indicate the specific slides which have been displayed in Figure 1.

It has been observed in the literature [12, 18], and can also be hinted at from Figure 2, that overdispersion is present in H2AX foci data so that, for instance, quasi-Poisson or negative binomial models appear adequate. The quasi-Poisson model is essentially a Poisson model which allows for variance/mean ratios different from one. A simple practical question arising is whether the model fitting can be carried out without loss of information using only the aggregated data, as displayed in Figure 2, or whether the raw data, as exemplified in Figure 1, should be used. This question is of greater depth than one would expect: while we demonstrate in the next section, that, in theory, one would anticipate the dispersion to be unaffected by the aggregation, for these data set the dispersion estimate resulting from a quasi-Poisson fit using the raw data is 1.223, while the one resulting from the fit to the aggregated data is 147.99! In a further twist, we will also see that the inflated dispersion of the aggregated model is not necessarily useless: it is a manifestation of a problem which lies elsewhere, namely dependency structures within the raw data, and eventually leads to the estimation of parameter standard errors which are more correct than those of the raw data model. Even though the connection of data aggregation to overdispersion is not an unknown phenomenon (in fact, in the ecological literature, the term 'aggregated' is often used synonymous to 'overdispersed' [2]), we believe that the implications of count data aggregation on dispersion estimates and ensuing inferential purposes, are, so far, poorly appreciated in the biodosimetric community, and also lack explicit study in the statistical literature.

The exposition is organised as follows. Section 2 summarises the statistical and conceptual basics of Poisson and quasi-Poisson models, including the estimation of the dispersion parameter for the latter, as well as its variance. Section 3 reports the results of the analysis of the above mentioned γ-H2AX data set as well as a bootstrap simulation (also including comparison to random effect models), which facilitates insight into the presence of heterogeneities as the source of inflation of the dispersion estimates, as well as the impact of this effect on standard errors of model parameters. Section 4 focuses on the special case of mixture-driven heterogeneity, deriving and validating through simulation the inflation of dispersion explicitly. Lastly, Section 5 concludes the paper, and discusses practical implications of our results. An Appendix which gives several addendums concerning data, code, derivations, and extensions, is also provided.

We close this introduction by outlining some notation. We refer to a set of foci counts (constituting a specific histogram such as in Figure 1) as a *slide*. The jth observation in the ith slide is denoted as y_{ij}, for k slides with respective size n_i, $i = 1, \dots, k$. Averaging over the ith slide, we obtain the means $y_i = \sum y_{ij}/n_i$, which are also referred to as *yields* in the dosimetry literature. The convention to speak of *slides* and *yields* is simply with reference to the data application considered, and is not implying a restriction of the validity of the results to this particular field of application. In particular, the section which follows is, statistically, written in a general context, albeit still using the terms laid out in this paragraph.

2 Raw and aggregated data models

2.1 Poisson models

The most basic count data model is the Poisson model, postulating

$$y_{ij} \sim \text{Pois}(\mu_i), \tag{2.1}$$

that is $f(y_{ij}|x_i) = \mathrm{e}^{-\mu_i}\mu_i^{y_{ij}}/y_{ij}!$, where

$$\mu_i = g^{-1}\left(x_i'\beta\right), \tag{2.2}$$

with $\mu_i > 0$, for some link function g. That is, observations corresponding to a particular slide share the same predictors, and hence the same μ_i. In some applications, the possible values of x_i (here, the design doses) may coincide with the grouping, or even define the grouping as such. Here, this is not the case, as we have multiple slides for a given dose. In either case, under this framework, the predictor x_i will never depend on j.

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

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

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

Another important characteristic of the Poisson model is that of *equidispersion*, that is, for all i and j,

$$\frac{\text{Var}(y_{ij}|x_i)}{E(y_{ij}|x_i)} = 1, \tag{2.3}$$

from which it is easy to see that also

$$\frac{\text{Var}(s_i|x_i)}{E(s_i|x_i)} = 1. \tag{2.4}$$

That is, the equidispersion carries over from the raw to the aggregated data model,

$$E(s_i|x_i) = n_i g^{-1}(x_i'\beta). \tag{2.5}$$

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

2.2 Overdispersed Poisson models

In practical data applications, the equidispersion property is frequently violated. In the most simple case, this violation can be described as generalization of (2.3),

$$\frac{\text{Var}(y_{ij}|x_i)}{E(y_{ij}|x_i)} = \phi \tag{2.6}$$

for some constant dispersion, $\phi > 0$. If $\phi > 1$ one speaks of overdispersion, while for $\phi < 1$ (which is less frequently encountered) one has underdispersion.

Poisson regression models can be easily adapted to allow for situation (2.6), since the dispersion cancels out from the score equations and so the estimates of regression parameters are unaffected. One speaks then of

quasi-Poisson regression models [19], which have gained interest specifically in the field of biodosimetry [12]. Under such a framework, standard errors can be conveniently computed in a post-hoc manner by multiplying the standard errors from the Poisson model with the square root of the (estimated) dispersion parameter [20].

While (2.6) is a considerable generalization of the 'plain' Poisson model, in practice the dispersion may depend on covariates x_i, that is $\phi = \phi(x_i)$. Covariate-dependent dispersion cannot be expressed by quasi-Poisson regression and requires more sophisticated models, such as the negative binomial [21, 22]. Such models are outside the scope of this paper. The relative advantages and disadvantages of quasi-Poisson and negative binomial models have been discussed in [23]. Further models for overdispersed data include the generalised Poisson [24], mixed Poisson [25], Hermite [26], and zero-inflated [27] models, the relationships among some of which are discussed in [28, 29].

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

$$\begin{aligned}\mathrm{Var}(s_i|x_i) = \mathrm{Var}\left(\sum_{j=1}^{n_i} y_{ij}|x_i\right) &\overset{*}{=} \sum_{j=1}^{n_i} \mathrm{Var}(y_{ij}|x_i) \\ &= \sum_{j=1}^{n_i} \phi E(y_{ij}|x_i) = \phi \sum_{j=1}^{n_i} E(y_{ij}|x_i) = \phi E(s_i|x_i),\end{aligned} \tag{2.7}$$

where the step $(*)$ is a consequence of conditional independence assumption; so once again $\mathrm{Var}(s_i|x_i)/E(s_i|x_i) = \phi$ so that the dispersion is, theoretically, invariant to aggregation.

2.3 Estimating dispersion

Typically when we speak of dispersion, we are referring to the variance divided by the mean. More precisely, following the notation in (2.6) and ignoring (for a moment) the presence of covariates, then dispersion is defined by

$$\phi = \frac{\mathrm{Var}(y_{ij})}{E(y_{ij})} \tag{2.8}$$

which can be estimated through the *dispersion index*

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

where $N = \sum_{i=1}^{k} n_i$ is the total number of observed counts, and $\bar{y} = N^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$ is their overall mean. Applying this on the full foci data set introduced in the Introduction provides us with a dispersion index of $\hat{\phi}_{\mathrm{ind}} = 1.494$.

Under the presence of covariates, which are related to the mean function μ_i via (2.2), the dispersion parameter ϕ can be estimated from the raw data model [20, 30] by

$$\hat{\phi}_{\mathrm{raw}} = \frac{1}{N-p} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(y_{ij} - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{2.9}$$

where p is the number of model parameters and $\hat{\mu}_i = g^{-1}(x_i' \hat{\boldsymbol{\beta}})$. McCullagh and Nelder [31] discuss the advantage of basing the estimation of the dispersion parameter on (2.9) as opposed to using the residual deviance.

For aggregated data $s_i = \sum_{j=1}^{n_i} y_{ij}$, $i = 1, \ldots, k$ (equivalently expressed through the yields $y_i = s_i/n_i$), with aggregated data model $E(s_i|x_i) = n_i g^{-1}(x_i'\beta)$, the value of the dispersion can be estimated by

$$\hat{\phi}_{\mathrm{agg}} = \frac{1}{k-p} \sum_{i=1}^{k} \frac{(s_i - n_i\hat{\mu}_i)^2}{n_i\hat{\mu}_i} = \frac{1}{k-p} \sum_{i=1}^{k} n_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, \tag{2.10}$$

where $\hat{\mu}_i$ is as above.

The standard errors associated with the estimated coefficients, $\hat{\boldsymbol{\beta}}$, will be the same as for the non-dispersed Poisson model inflated by the factor $\hat{\phi}^{1/2}$.

2.4 Variance of dispersion

A possible source of increased dispersion for the aggregated data model as compared to the raw data model could be an increased variance of the estimates under the former. In this subsection, we therefore study the effect of data aggregation on that variance. For this purpose, let us assume that the true dispersion, ϕ, is indeed the same for the raw and aggregated data. We then note that the formulations presented in (2.9) and (2.10) are the respective Pearson X^2 goodness-of-fit statistics divided by the residual degrees of freedom, which we denote as ν, such that $\hat{\phi}_\nu = X^2/\nu$. Specifically, we have $\nu = N - p$ for the raw data and $\nu = k - p$ for the aggregated data, so $\hat{\phi}_{\text{raw}} = \hat{\phi}_{N-p}$ and $\hat{\phi}_{\text{agg}} = \hat{\phi}_{k-p}$.

If indeed our fitted model is correct then we would expect X^2/ϕ to have a χ^2_ν distribution [20], implying that $E(X^2) = \phi\nu$ and $\text{Var}(X^2) = 2\phi^2\nu$. It then follows that

$$E(\hat{\phi}_\nu) = \frac{1}{\nu}E(X^2) = \phi, \tag{2.11}$$

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

$$\text{Var}\left(\hat{\phi}_\nu\right) = \frac{1}{\nu^2}\,\text{Var}\left(X^2\right) = \frac{2\phi^2}{\nu}. \tag{2.12}$$

It is then clear that for the aggregated model, where ν is much smaller, $\text{Var}(\hat{\phi})$ is larger.

3 Application to H2AX foci data

We now apply the methodology outlined in the previous section to the γ-H2AX foci dataset presented in the Introduction.

3.1 Fitting raw and aggregated data models

In the context of γ-H2AX foci analysis, it is common to consider a linear model for foci counts with identity link $g(\mu) = \mu$ [12], such that the raw data model becomes

$$\mu_i = E(y_{ij}|x_i) = \alpha + \beta x_i, \tag{3.1}$$

with corresponding aggregated data model

$$E(s_i|x_i) = \alpha \times n_i + \beta \times (n_i x_i), \tag{3.2}$$

where x_i now denotes dose. This choice is motivated by physical considerations and the shape of the dose-response curve, despite the fact that the log-link is, from a statistical viewpoint, a natural choice for count data.

We now fit raw and aggregated data models (3.1) and (3.2) to the previously introduced H2AX foci data. Dispersions are estimated via (2.9) and (2.10). The results from both model fits are presented in Table 1. We note that the coefficients of the quasi-Poisson models do not change between the two data types. Hence, the calibration curves of expected foci yield given dose, as displayed in Figure 3, will remain exactly the same if estimated through raw or aggregated data models. However, a significant difference is observed in their dispersions, where for our data we obtain estimates of $\hat{\phi}_{\text{raw}} = 1.223$ and $\hat{\phi}_{\text{agg}} = 147.99$. It is noted that both of these would lead to a rejection of the Poisson hypothesis with a χ^2 goodness of-fit-test (see Table 1).

Table 1: Parameter estimates along with their associated standard errors and dispersion estimates obtained from each model.

	Raw	Aggregated
$(\hat{\alpha}, \hat{\beta})$	*(2.011, 5.746)*	*(2.011, 5.746)*
$(SE(\hat{\alpha}), SE(\hat{\beta}))$	*(0.009, 0.023)*	*(0.102, 0.248)*
$\hat{\phi}$	*1.223*	*147.99*
SE$[\hat{\phi}]$	0.004	0.16
ν	233218	114
$\chi^2_{\nu,0.95}/\nu$	1.005	1.227

Results shown in italics stem directly from the fitted models; the standard errors of $\hat{\phi}$ and the values of ν follow the rationale set out in Section 3.1. The last row gives the critical value that $\hat{\phi}$ would be compared with in a Poisson goodness-of-fit test at the 5% level of significance.

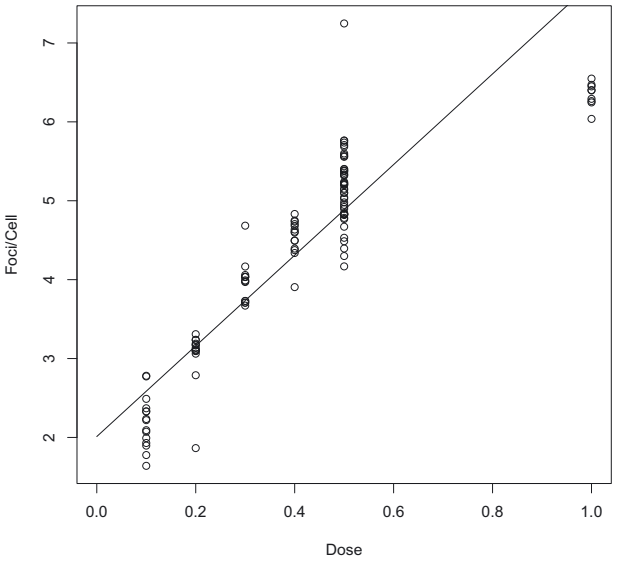

Figure 3: Quasi-Poisson model estimates of the linear calibration curve: $E(y_i) = 2.011 + 5.746x_i$.

We proceed with investigating whether an increased variance of dispersion, as discussed in Section 2.4, could be responsible for this effect. Specifically, we formulate this question as follows: assuming that the value $\hat{\phi}_{\text{raw}} = 1.223$ represents the true dispersion of the models, then is it possible *to obtain aggregated dispersion estimates of the magnitude of 150 purely due to increased variance*? Substituting $\hat{\phi}_{\text{raw}} = 1.223$ in the right hand side of (2.12), with degrees of freedom adjusted according to Table 1, leads to $\text{SE}(\hat{\phi}_{\text{raw}}) \approx 0.004$ for the raw models and $\text{SE}(\hat{\phi}_{\text{agg}}) \approx 0.16$ for the aggregated models. However, this effect – to which we refer as *variance effect* henceforth – is certainly not sufficient to explain a value of, say, $\hat{\phi}_{\text{agg}} = 147.99$, for the dispersion of the linear fit to the aggregated data.

3.2 Random effect models

Since the data y_{ij} do possess a two-level structure, with the slides i corresponding to the upper level, and the foci frequencies within slides corresponding to the lower level, it appears adequate to contrast the previous results with an alternative modelling strategy where within-slide correlation is explicitly accounted for by an additive random effect, also called random intercept, operating on the upper level. Hence, we consider a mean function of type $\tilde{\mu}_i = \mu_i + u_i$, where μ_i is as in (3.1), and $u_i \sim N(0, \sigma_r^2)$ is a Gaussian random effect. For the response distribution, we consider two scenarios, namely a Poisson mixed model $y_{ij} \sim \text{Pois}(\tilde{\mu}_i)$, and an

Table 2: Parameter estimates of the fitted random effect models.

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

Results shown in italics are extracted directly from the output of function `glmmTMB`. The values below the dashed line give the estimated residual variance, $\hat{\sigma}_\epsilon^2$, and the resulting ICC values.

NB (Type 1) regression model, $y_{ij} \sim \text{NB}(\tilde{\mu}_i, a)$ where $\phi = 1 + a$. That is, the NB1 model allows the parameter ϕ to capture any dispersion not accounted for by the slide-wise random effect.

The models are fitted with R function `glmmTMB` [32], and results are provided in Table 2. We firstly observe that both models behave similarly, and that their standard errors lend, interestingly, support to the aggregated data model. This can be interpreted as that the dispersion estimate of the aggregated model has successfully captured the between-slide heterogeneity described by the random effect model. Informally, the presence of this heterogeneity is visible from the small but non-zero intra-class correlations (ICC) in Table 2. More formally, one can carry out statistical tests for the significance of the random effect term, with $H_0: \sigma_r^2 = 0$. For the Poisson model, the likelihood ratio statistic of models with and without the random effect term is $2(513,385.3 - 505,138.5) = 16,493.6$, clearly indicating rejection of H_0 when contrasting with a $0.5(\chi_0^2 + \chi_1^2)$ distribution. For the NB1 model, the conclusion is identical with LR $= 2(511119.5 - 503979.3) = 14280.4$. One can test for the significance of overdispersion ($H_0: \phi = 1$) by comparing $\hat{\phi} = 1.141$ with $\chi^2_{0.95,233,217}/233,217 = 1.005$, also yielding significance. So, albeit just above 1, the value of 1.141 represents genuine overdispersion (over and above the one explained by the random effect model). In summary, this provides evidence of heterogeneities existing both between and within slides. It is furthermore noted that the coefficient estimates of $\hat{\alpha}$ and $\hat{\beta}$ for the random effect model differ by about three standard errors from the raw and aggregated data models.

3.3 Bootstrap simulation

Having seen the evidence for heterogeneities in the data, the models fitted in Section 3.1 can be considered misspecified. In order to understand better the impact of this misspecification on the fitted raw and aggregated data models, we carry out a bootstrap simulation, with the mixed NB1 model fitted in Section 3.2 as base model, and examine the dispersion estimates, and resulting standard errors, of all models.

The sampling process of this bootstrap is built in two stages (with all estimates taken from Table 2):

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

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

(i) A quasi-Poisson regression model with identity link, applied on the bootstrapped raw data y_{ij}^*, i.e. model (3.1).
(ii) A quasi-Poisson regression model with identity link, applied on the bootstrapped aggregated data $s_i^* = \sum_j y_{ij}^*$ i.e. model (3.2).
(iii) A NB1 regression model with identity link, applied on the bootstrapped raw data y_{ij}^*; with an additive random effect representing slides.

For each fitted model and bootstrap iteration, dispersion estimates for models (i) and (ii) are computed according to (2.9) and (2.10), respectively, with standard errors arising as explained at the end of Section 2.3.

For model (iii), this dispersion estimate is obtained by adding 1 to the 'overdispersion' parameter, $\hat{a}$, reported in the summary output of R function `glmmTMB` [32]. Standard errors are extracted directly from this output.

Boxplots of the dispersion estimates for the bootstrap simulation are displayed in Figure 4. The left hand panel in this figure gives a comparison of the dispersion estimates for the raw, random, and aggregated models, whereas the right panel gives a zoomed comparison of the raw and random effect models. We see from this that the dispersion estimates for the raw data model are positioned close to the correct mean value at 1.223. However, the boxplot for the dispersion estimates from the aggregated model now sits at about 160, which is of similar magnitude as in our initial analysis displayed in Table 1. While the variability of these estimates is also larger than for the raw data model, it is clear that something much more drastic (than just inflation of variance) has occurred here, shifting the bulk of the dispersion estimates from the magnitude 1–2 to much larger values. The dispersion estimates from the random effect model are slightly smaller than for the raw data model, centering correctly at the value 1.141 from which the data were generated, as visible from the right panel. The slight difference between these two models is plausible, as some of the original overdispersion has been captured by the random effect.

We investigate now the consequences of this inflated dispersion. Therefore, let us firstly consider the boxplots in Figure 5. It is clear from this that, for the raw data model, the reported standard errors of $\hat{\alpha}$ and $\hat{\beta}$ are very small. However, either of aggregation, or the use of a random effect, transports the standard errors to

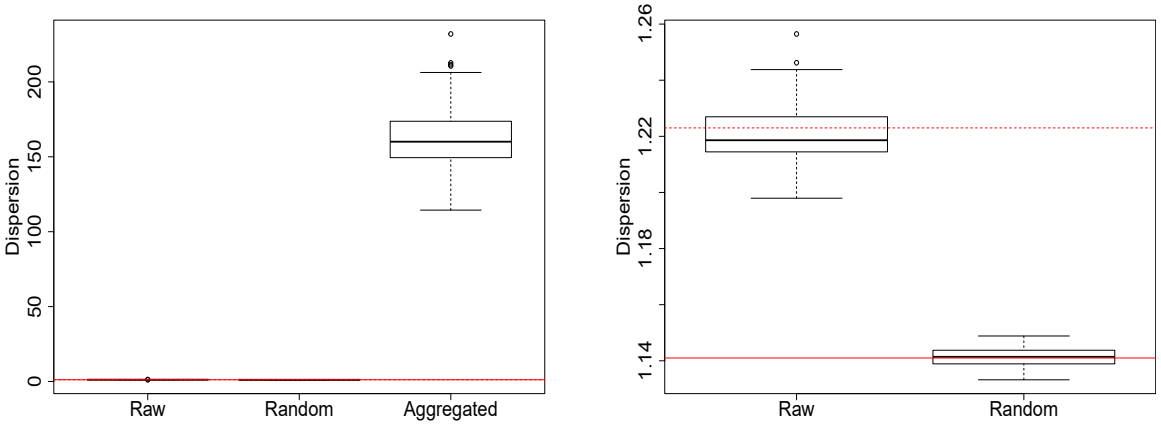

Figure 4: Dispersion estimates based on the bootstrap simulation. The solid red line represents the random-effect model dispersion $\hat{\phi} = 1.141$ and the dashed line indicates the quasi-Poisson dispersion $\hat{\phi} = 1.223$ for the original data as reported in Tables 1 and 2.

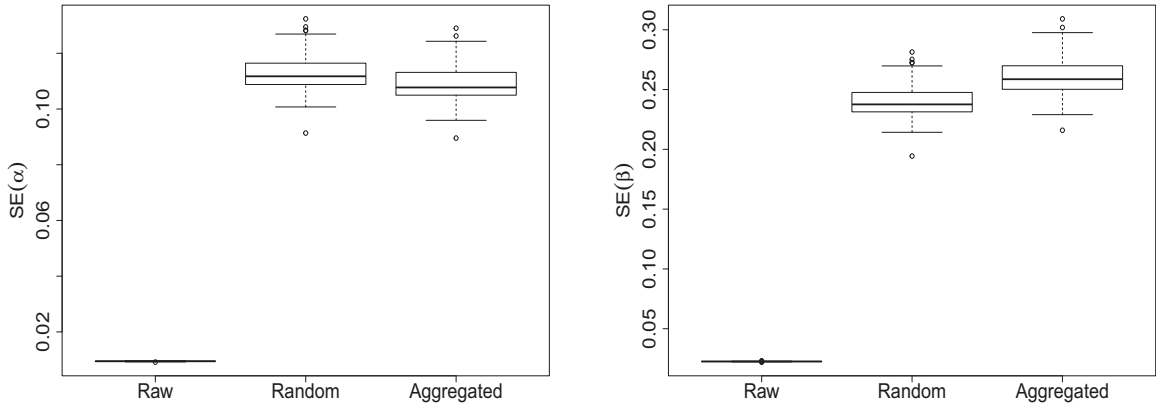

Figure 5: Parameter standard errors for the bootstrap simulation (left: intercept; right: slope).

Table 3: Parameter standard deviations based on 100 simulation runs.

	Raw	Random	Aggregated
SD($\hat{\alpha}$)	0.109	0.103	0.109
SD($\hat{\beta}$)	0.226	0.211	0.226

a much higher level, as also displayed in Table 1. The values reported in Table 3 reveal that, over all estimation methods, the actual *standard deviation* of the bootstrapped estimates of regression coefficients is very similar, and is *for all three models, including the raw data model*, of the (high) magnitude reported by the aggregated and random effect models. This, in turn, implies that the standard errors of regression parameters for the raw data model, as reported in Table 1 and Figure 5, are wrong. We arrive, hence, at the intriguing conclusion that the large dispersion produced by the aggregated data model serves eventually a good purpose — namely to adjust the standard errors of the parameter estimates so that these match the magnitude of those from the random effect model. For later reference, we will refer to this effect, i.e. the tendency of aggregated data models to inflate dispersion estimates in order to account for violations of the independence assumption in the raw data, as a *dependency effect*.

4 Special case: mixture-induced heterogeneity

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

4.1 A two-component model inducing heterogeneity

Consider a scenario in which we generate k rows (slides), each consisting of $n_i \equiv n$ Poisson foci counts (cells), but for fixed covariate dose (in other words, in the absence of covariates). However, we assume that there exists heterogeneity, that is some counts are from a Pois(λ_1) distribution with probability q (the Bernoulli parameter which selects the Poisson mean) and others from a Pois(λ_2) with probability $1-q$. In general terms, the Poisson means come from a two-point mixture; i.e. each raw count y_{ij} is generated as

$$y_{ij} \sim Z_{ij}\text{Pois}(\lambda_1) + (1 - Z_{ij})\text{Pois}(\lambda_2) \tag{4.1}$$

where $Z_{ij} \sim B(1, q)$. The resulting heterogeneity creates overdispersion which, under model (4.1), can be exactly quantified as

$$\phi = \frac{\text{Var}(y_{ij})}{E(y_{ij})} = 1 + \frac{q(1-q)(\lambda_1 - \lambda_2)^2}{q\lambda_1 + (1-q)\lambda_2}. \tag{4.2}$$

See Appendix B.2 for proof of this statement and Figure 6 (top) for a visual representation of ϕ as a function of q; note also that the dependence on x_i as in (2.6) is now suppressed as there are no covariates. Expression (4.2) holds true even if there are correlation structures within the Z_{ij}. However, we will see that, for the dispersion of the aggregated data, it makes a crucial difference whether the heterogeneity is entirely random (i.e. the indicators Z_{ij} are independently generated for all i and j), or whether there is some correlation structure.

Consider, for instance, a scenario in which

$$Z_{ij} \equiv Z_i \quad \text{for all} \quad j = 1, \dots, n, \tag{4.3}$$

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

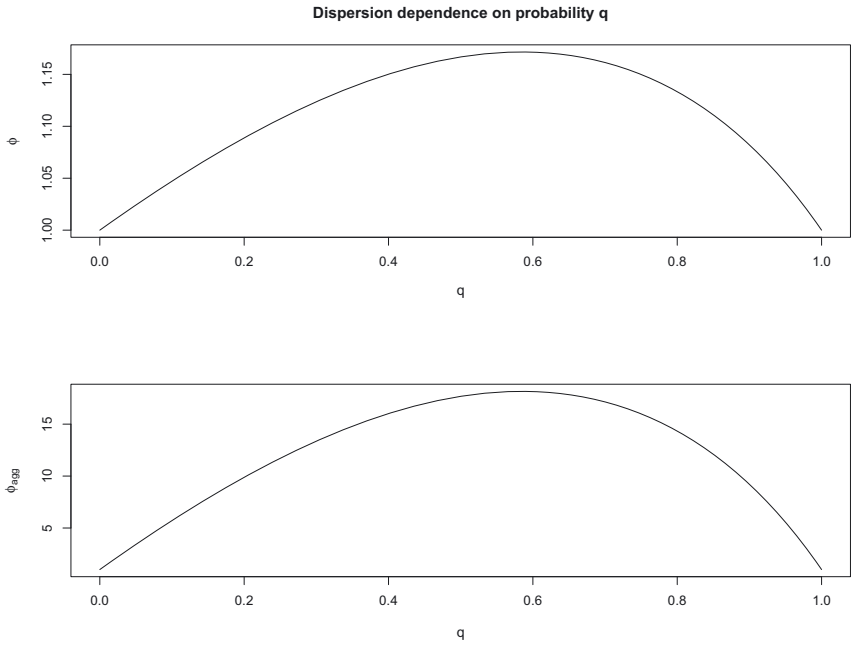

Figure 6: For fixed $\lambda_1 = 1$, $\lambda_2 = 2$, we plot the non-linear functions (4.2) and (4.9), using a string size of $\tau = 100$. Note the substantially different scales in the vertical axes of the two plots.

$$\begin{aligned}\mathrm{Cov}(y_{ij}, y_{il}) &= E(\mathrm{Cov}(y_{ij}, y_{il})|Z_i) + \mathrm{Cov}(E(y_{ij}|Z_i), E(y_{il}|Z_i)) \\ &= \lambda_1^2 \mathrm{Var}(Z_i) + \lambda_2^2 \mathrm{Var}(1 - Z_i) + 2\lambda_1\lambda_2 \mathrm{Cov}(Z_i, 1 - Z_i) \\ &= q(1-q)(\lambda_1 - \lambda_2)^2 \end{aligned} \tag{4.4}$$

so for $\lambda_1 \neq \lambda_2$ the independence assumption in (*) in Section 2 is clearly violated. Depending on the mechanism generating the Z_{ij}, this expression will look different, but the point is that any dependency structures within the Z_{ij} will render these covariances non-zero.

4.2 Theoretical dispersion of aggregated data

Aggregated data are obtained as before as $s_i = \sum_{j=1}^{n} y_{ij}$. The object of interest in this subsection is $\phi_{\mathrm{agg}} = \mathrm{Var}(s_i)/E(s_i)$, where we have now made notationally explicit that it may be different from ϕ. Through the law of total expectation and variance one can show that (see Appendix B.2), under model (4.1)

$$E(s_i) = n(q\lambda_1 + (1-q)\lambda_2); \tag{4.5}$$

$$\mathrm{Var}(s_i) = n(q\lambda_1 + (1-q)\lambda_2) + nq(1-q)(\lambda_1 - \lambda_2)^2 + \sum_{j \neq l}^{n} \mathrm{Cov}(y_{ij}, y_{il}). \tag{4.6}$$

This gives a general expression for the aggregated dispersion,

$$\begin{aligned}\phi_{\mathrm{agg}} &\equiv \frac{\mathrm{Var}(s_i)}{E(s_i)} \\ &= 1 + \frac{q(1-q)(\lambda_1 - \lambda_2)^2}{q\lambda_1 + (1-q)\lambda_2} + \frac{\sum_{j \neq l}^{n} \mathrm{Cov}(y_{ij}, y_{il})}{n(q\lambda_1 + (1-q)\lambda_2)}. \end{aligned} \tag{4.7}$$

In the simplest case that all covariances are identical to 0, the third term disappears and one sees immediately that ϕ_{agg} corresponds to the expression for ϕ given in (4.2). In the previously discussed case of slide-wise dependencies (4.3), one finds by using expression (4.4) and then referring to (4.2) that

$$\phi_{\text{agg}} = 1 + \frac{nq(1-q)(\lambda_1 - \lambda_2)^2}{q\lambda_1 + (1-q)\lambda_2} = 1 + n\,(\phi - 1)\,. \tag{4.8}$$

We discuss a third scenario which we consider of practical relevance. Assume there are correlated strings of length $\tau < n$, each sharing the same Poisson mean. One can consider this as a special case of model (4.1) where the indicators Z_{ij} share the same value for blocks of length $\tau < n$, in terms of the index j. Then one can show (Appendix B.2) that

$$\phi_{\text{agg}} = 1 + \tau\,(\phi - 1)\,, \tag{4.9}$$

neatly extending (4.8). Note that both ϕ and ϕ_{agg} can be considered as functions of the mixing proportion, q. This is visualised in Figure 6 (bottom). We take note of the non-symmetry in terms of the mixing parameter, with a maximum at $q = 2 - \sqrt{2}$. Furthermore, we observe that for $q = 0$ or $q = 1$ there is no overdispersion since there is no heterogeneity.

Equations (4.8) and (4.9) provide some insight into how the presence of different types of heterogeneity, for example through correlation within rows or strings within rows, affect the dispersion of the aggregated data. From direct inspection of both (4.8) and (4.9), we deduce that if one increases either the row length or the string size then the dispersion of the aggregated data continues to grow larger. We also notice that if there is no overdispersion of the raw counts, i.e. $\phi = 1$, then we have equidispersion for the aggregated data as expected. If one has only clusters of size 1 ($\tau = 1$ or $n = 1$; that is, the heterogeneity is entirely random) then $\phi_{\text{agg}} = \phi$, so in this case the aggregated data dispersion does not inflate.

4.3 Experiment

We carry out a simulation experiment as described in Section 4.1 using $\lambda_1 = 1$, $\lambda_2 = 2$ and $q = 0.5$. The mechanisms presented in Section 4.1 and the theoretical derivations in Section 4.2 mean that the heterogeneity resulting from the mixture will trigger overdispersion, but that the overdispersion for the aggregated data will depend on the correlation structure of the heterogeneity-inducing mechanism. This leads us to distinguish the following three cases:

(A) Random heterogeneity: for each slide and cell, the Z_{ij} in (4.1) are generated independently;
(B) Slide-wise heterogeneity: the Z_{ij} are generated once for each slide and kept constant for all cells in that slide, i.e. $Z_{ij} = Z_i$ as in (4.3);
(C) String-wise heterogeneity: the Z_{ij} share the same value for blocks of size $\tau = 100$ within each slide, but different blocks are generated independently.

For each of (A), (B) and (C), $k = 1000$ slides of length $n = 1000$ are generated. Since no covariates are involved in this study, we do not need to fit any models to estimate dispersion. For the raw data, the dispersion is estimated by the overall dispersion index (2.8). For the aggregated data, this would be replaced by $\sum_{i=1}^{k}(s_i - \bar{s})^2/[(k-1)\bar{s}]$. The resulting dispersion values are reported in Table 4, with corresponding R code detailed in Appendix B.1.

Table 4: Dispersion indexes from simulated data under scenarios (A), (B) and (C).

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

We can see that for case (A) the dispersion of the aggregated data does not increase at all, while in (B) we observe the strongest inflation. To reiterate, 'aggregated data' signifies here row-wise (slide-wise) sums. Our γ-H2AX data best corresponds to (C) rather than (B), although the basis of the effect is the same.

Verifying these results through our theoretical derivations from Section 4.2, one obtains for case (B) via (4.8) that

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

which agrees closely with the simulated value of 168.97. Under scenario (C), where slides are split into 10 clusters each containing 100 cells ($\tau = 100$), one gets from (4.9)

$$\phi_{\text{agg}} = 1 + 100\,(1.166 - 1) = 16.66,$$

again in good agreement with our simulation result of 16.85.

4.4 Generalization of the model

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

As the Multinomial model is often not suitable when there is observed over-dispersion, the Dirichlet multinomial distribution model can be used as an alternative [33]. For further consideration, one may consider a model of type $y_{ij} = \sum_m Z_{\text{ijm}}\text{Pois}(\lambda_m)$ where variations among the component probabilities $q_m = P(Z_{\text{ijm}} = 1)$ follow a Dirichlet distribution, i.e. $q_m \sim \text{Dir}(\alpha)$, $m = 1, \ldots, M$, indicating that y_{ij} belongs to component m with probability q_m.

5 Discussion

In many applied sciences, the use of aggregated count data is common, since they contain all relevant information to estimate Poisson models. Aggregated data are also usually less expensive to store and analyse than individual data. For instance, in biodosimetry, aggregated data are popularly used for many biomarkers including the dicentric chromosome assay [34, 35]. Another reason for the use of aggregated data is just convenience: while for biomarkers based on chromosomal aberrations, such as the dicentric assay, where counts larger than 7 or 8 are rarely observed, the full count distributions can still be conveniently displayed [36], this is not necessarily the case for H2AX foci data where this count may be much higher. The data analyst may never get to see the raw data, and then has to work with the aggregated data simply as this is all that is available to them [12].

The early literature on the γ-H2AX assay reported that the Poisson assumption is well fulfilled for manually scored H2AX foci data, and also provided a biological argument relating to the random induction of double-strand breaks which underlines this point [8, 18]. However, even under manual scoring, deviations from this property can pervade through multiple sources of heterogeneity, leading to overdispersion [12]. For automatic foci counting [37], this is exacerbated as it introduces additional variability and technical artefacts [18].

Under the presence of overdispersion, a conditional independence assumption of the responses given covariates guarantees, in theory, equality of the raw and aggregated data dispersion. However, we have seen that dispersion estimates for raw and aggregated data can differ dramatically for practical data sets. We distinguished that there are two effects which jointly result in an increased dispersion for the aggregated data model; a (relatively minor, but still significant) variance effect and a (potentially huge) dependency effect. We have demonstrated the latter phenomenon via example, simulation, and theory, uncovering in this process that the causes for the dependency effect reside in correlations between or within the slides being

aggregated over. Another way of putting these findings is: the presence of unobserved heterogeneity will cause overdispersion in the raw data. If this heterogeneity follows dependency patterns (within or between slides), then this will lead to *inflated* overdispersion for the aggregated data.

Several experimental factors may contribute to overdispersion in the raw data. A certain role is played by technical variations, such as in the intensity filter used for the foci scoring. Specifically, for low foci rates the semi-automated imaging software which aids the foci scoring tends to produce spurious foci by over-enhancing background signals. Other sources of overdispersion may relate to physical issues with the slides, issues with the radiation source itself or the placement of the samples, issues relating to the antibodies used to produce foci, the microscope, and the scorer. Any of these issues may also incur dependencies, for instance it is likely to assume a 'learning effect' for the scorer who may be tempted to discard samples which do not fit the previously observed pattern. While the theoretical derivations, in Section 4, only cover the covariate-free case, they still give useful insights into the relationship of raw and aggregated data dispersion; specifically the aggregated dispersion increases linearly with the length of correlated strings within the data set, attaining a maximum if the string size corresponds to the full slides.

The relevant question is then whether the raw data should have been used if they were available, and if so, using which model. Under the presence of, say, slide-wise correlations in the raw data, the statistically sound model would be the use of a mixed model for the raw data which features a random intercept for each slide. It appears that such a model produces roughly similar parameter standard errors than the aggregated data model, whereas the raw data model produces much smaller standard errors. This appears to indicate that the high dispersion produced by the aggregated model is an attempt by the model to solve a problem which resides somewhere else (namely in the between-slide-correlations), which the raw data model is not able to address (without the inclusion of random effects). Putting it into other words, the aggregated model finds a way to produce roughly *correct* uncertainty quantification by using *incorrect* dispersion estimates.

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

A practical advice to laboratories is to reduce heterogeneities to an absolute minimum, as they inflate dispersions and standard errors, and may also shift the actual calibration curve parameters. We do advise against using raw data models without adjustment by a random effect, but we do not advise against using the aggregated data models. Aggregation on the slide level does account for the correlations just as the random effect model would do, albeit using a much simpler model. However the data analyst should be aware that the resulting dispersion estimates may be far from the underlying true dispersion of the raw data. This is of particular importance with view to the detection of partial body exposures through dispersion estimates, as is a common approach for dicentric chromosomes [35]. Inflated dispersions of the magnitude as observed in this paper would certainly render any attempt at identifying partial body exposure ineffective, unless one finds a way of working backwards to recover the raw data dispersion, for instance using equations such as derived in Section 4.2. This question is left for future investigation.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: This work has arisen from activities which have been partially supported by the Durham Research Impact Fund (RIF).
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

Appendix

A Data generation and cleaning

Blood was collected from healthy donors via an 18- or 20-gauge indwelling cannula (Vasofix Safety IV; B. Braun Melsungen AG, Melsungen, Germany) into 7.5 ml lithium heparin monovettes (S-Monovette; Sarstedt

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

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

Search and image acquisition of cell nuclei on the slides was performed by automatic fluorescence microscopy using a scanning and imaging platform (Metafer 4, version V3.13.1; Meta-Systems Hard & Software GmbH, Altlussheim, Germany) equipped with an objective (ZeissPlan-Neofluar 40× 0.75; Carl Zeiss Microscopy GmbH, Jena, Germany) yielding a 400-fold magnification. For foci analysis a Spectrum Orange bandpass filter (excitation: center wavelength/bandwidth = 546/10 nm, emission: 580/30 nm; Chroma 31003; Chroma Technology, Olching, Germany) and for counterstaining a DAPI bandpass filter (excitation: 350/50 nm, emission: 460/50 nm; Chroma 31000; Chroma Technology) was used. A foci specific Classifier 2.0.1 was created and used in all experiments.

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

B Violation of quasi-Poisson independence

B.1 Simulation

In Section 4, we verified through simulation the dependency effect (for a fixed covariate value dose) through three different heterogeneity cases. The R code to reproduce the results in Table 4 is presented below.

```
intercepts <- c(1,2) # these are the two possible Poisson means
# lambda_1 and lambda_2
q <- c(0.5, 0.5) # probability q = 0.5
jmax = 1000
j<-1
yM <- matrix(0, 1000, jmax)
while (j <=jmax){
  # Run one of the following three commands:
  # (A) all Poisson means are independently chosen
  # r.intercepts <- sample(intercepts, 1000, replace=TRUE, prob=q)
  # (B) all Poisson means are the same for a fixed row
  # r.intercepts <- sample(intercepts, 1, replace=TRUE, prob=q)
  # (C) within each row, strings of size 10 share the same
  # mean
  # r.intercepts <- rep(sample(intercepts, 10, replace=TRUE, prob=q),
  # each=100)
  xM <-rep(0,1000) # dose = 0
  yM[,j]<- rpois(1000, r.intercepts) # generates Poisson counts
  j<-j+1
  if ((j %%10) ==0){print(j)}
}
# (A)
var(as.vector(yM))/mean(as.vector(yM)) # raw dispersion
# [1] 1.16772
var(colSums(yM))/mean(colSums(yM)) # aggregated dispersion
# [1] 1.070203

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

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

B.2 Theoretical derivation

We now present the theory behind the dispersion estimates for the two-component mixture model (4.1). We begin with deriving (4.5) and (4.6). Recall that y_{ij} denotes the jth count (cell) for slide i with $j = 1, \dots, n$, and that $Z_{ij} \sim B(1, q)$, where yet no assumptions on the dependency structure of the Z_{ij} are being made. Then,

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

and

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

By dividing these two expressions, the dispersion index for the individual counts becomes (4.2). Now consider aggregated counts $s_i = \sum_{j=1}^{n} y_{ij}$. Then

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

and

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

which after division gives (4.7).

Consider now the special case $Z_{ij} \equiv Z_i$ (4.3). Then from (4.7) and (4.4),

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

which proves (4.8).

Now assume the slide with n cells consists of $b = \frac{n}{\tau}$ sub-groups (or strings) of size τ, where all y_{ij} in each batch are generated from the same distribution (either $\mathrm{Pois}(\lambda_1)$ with probability q or $\mathrm{Pois}(\lambda_2)$ with probability $1-q$). (In terms of the experiment in Section 4.3, this setup corresponds to scenario (C) but covers scenario (B) in the case $\tau = n$, and scenario (A) in the case $\tau = 1$). This general model is hence formulated as

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

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

The only required modification as compared to the previous derivation is to work out the covariances in the third term of (4.6). Observe here that the result (4.4) remains true but only for the observations within each string, that is

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

This implies

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

so that

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

Hence,

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

which is just (4.9).

B.3 3-Component Poisson mixture

Assume that some observations are from a Pois(λ_1) distribution with probability q_1, some from a Pois(λ_2) with probability q_2 while others are from a Pois(λ_3) with probability $q_3 = 1-q_1-q_2$. Each raw count y_{ij} is generated as

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

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

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

For the aggregated data, defined before as $s_i = \sum_{j=1}^{n} y_{ij}$, one arrives at the same expressions as in (4.8) and (4.9). The resulting dispersion values corresponding to the three heterogeneity scenarios with $\lambda_1 = 1$, $\lambda_2 = 2$ and $\lambda_3 = 3$ and equal probabilities i.e. $q_1 = q_2 = q_3 = 1/3$ are reported above in Table 5.

For comparison, under case (B) with $n = 1000$ one obtains from (4.8) that

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

and for scenario (C) with $\tau = 100$ in (4.9)

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

therefore in fairly good agreement.

Table 5: Dispersion indexes from simulated data under simulation scenarios as outlined in Section 4.3.

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

References

1. Hasnine, MN. Towards final scores prediction over clickstream using machine learning methods. In: Asia-Pacific society for computers in education (APSCE) 2018.
2. Harrison XA. Using observation-level random effects to model overdispersion in count data in ecology and evolution. PeerJ 2014;2:e616.
3. Rogakou E, Pilch D, Orr A, Ivanova V, Bonner W. Dna double-stranded breaks induce histone H2AX phosphorylation on serine 139. J Biol Chem 1998;273:5858–68.
4. Kuo L, Yang L. Gamma-H2AX – a novel biomarker for dna double-strand breaks. In Vivo 2008;22:305–9.
5. Barnard S, Bouffler S, Rothkamm K. The shape of the radiation dose response for dna double-strand break induction and repair. Genome Integr 2013;4:1.
6. Mandina T, Roch-Lefèvre S, Voisin P, González J, Lamadrid A, Romero I, et al. Dose-response relationship of gamma-H2AX foci induction in human lymphocytes after x-rays exposure. Radiat Meas 2011;46:997–9.
7. Roch-Lefèvre S, Mandina T, Voisin P, Gaëtan G, Mesa J, Valente M, et al. Quantification of gamma-h2ax foci in human lymphocytes: a method for biological dosimetry after ionizing radiation exposure. Radiat Res 2010;174:185–94.
8. Rothkamm K, Horn S. Gamma-H2AX as protein biomarker for radiation exposure. Ann Ist Super Sanita 2009;45:265–71.
9. Rothkamm K, Horn S, Scherthan H, Rossler U, De Amicis A, Barnard S, et al. Laboratory intercomparison on the gamma-h2ax foci assay. Radiat Res 2013;180:149–55.
10. Ainsbury E, Higueras M, Puig P, Einbeck J, Samaga D, Barquinero J, et al. Uncertainty of fast biological radiation dose assessment for emergency response scenarios. Int J Radiat Biol 2017;93:127–35.
11. Ainsbury EA, Samaga D, Della Monaca S, Marrale M, Bassinet C, Burbidge CI, et al. Uncertainty on radiation doses estimated by biological and retrospective physical methods. Radiat Protect Dosim 2017;178:382–404.
12. Einbeck J, Ainsbury E, Sales R, Barnard S, Kaestle F, Higueras M. A statistical framework for radiation dose estimation with uncertainty quantification from the γ-H2AX assay. PloS One 2018;13:e0207464.
13. Kopp B, Khoury L, Audebert M. Validation of the γh2ax biomarker for genotoxicity assessment: a review. Arch Toxicol 2019;93:2103–14.
14. Khoury L, Zalko D, Audebert M. Evaluation of the genotoxic potential of apoptosis inducers with the γh2ax assay in human cells. Mutat Res Genet Toxicol Environ Mutagen 2020;852:1–10.
15. Redon C, Nakamura A, Martin O, Parekh P, Weyemi U, Bonner W. Recent developments in the use of γ-h2ax as a quantitative dna double-strand break biomarker. Aging 2011;3:168–74.
16. Brix G, Gunther E, Rossler U, Endesfelder D, Kamp A, Beer A, et al. Double-strand breaks in lymphocyte dna of humans exposed to [18f] fluorodeoxyglucose and the static magnetic field in pet/mri. Eur J Nucl Med Mol Imag 2020;10:1–11.
17. Moquet J, Barnard S, Staynova A, Lindholm C, Monteiro Gil O, Martins V, et al. The second gamma-h2ax assay inter-comparison exercise carried out in the framework of the european biodosimetry network (reneb). Int J Radiat Biol 2017;93:58–64.
18. Rothkamm K, Barnard S, Ainsbury E, Al-Hafidh J, Barquinero J, Lindholm C, 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? Mutat Res 2013;756:170–73.
19. Wedderburn R. Quasi-likelihood functions, generalized linear models, and the gaussnewton method. Biometrika 1974;61:439–47.
20. Fahrmeir L, Tutz G. Multivariate statistical modelling based on generalized linear models. New York: Springer; 2011.
21. Cameron C, Trivedi P. Econometric models based on count data: comparisons and applications of some estimators and tests. J Appl Econom 1986;1:1.
22. Lloyd J. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PloS One 2007;2:e180.
23. Ver Hoef J, Boveng P. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology 2007;88:2766–72.
24. Consul P. Generalized Poisson distributions: properties and applications. New York: Marcel Dekker; 1989.
25. Wang P, Puterman M, Cockburn I, Le N. Mixed Poisson regression models with covariate dependent rates. Biometrics 1996;52:381–400.
26. Puig P, Barquinero J. An application of compound Poisson modelling to biological dosimetry. Proc R Soc A 2011;467:897–910.
27. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 1992;34:1–14.
28. Joe H, Zhu R. Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. Biom J 2005;47:219–29.
29. Lord D, Washington S, Ivan J. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid Anal Prev 2005;37:35–46.

30. Hinde J, Demétrio C. Overdispersion: models and estimation. Comput Stat Data Anal 1998;27:151–70.
31. McCullagh P, Nelder J. Generalized linear models. Chapman and Hall/CRC monographs on statistics and applied probability series, 2nd ed. London: Chapman & Hall; 1989.
32. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J 2017;9:378–400.
33. Joe H, Zhu R. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika 1962;49:65–82.
34. Ainsbury EA, Lloyd DC. Dose estimation software for radiation biodosimetry. Health Phys 2010;98:290–5.
35. Hilali A, Léonard E, Decat G, Léonard A. An appraisal of the value of the contaminated Poisson method to estimate the dose inhomogeneity in simulated partial-body exposure. Radiat Res 1991;128:108–11.
36. Oliveira M, Einbeck J, Higueras M, Ainsbury E, Puig P, Rothkamm K. Zero-inflated regression models for radiation-induced chromosome aberration data: a comparative study. Biom J 2016;58:259–79.
37. Ivashkevich AN, Martin OA, Smith AJ, Redon CE, Bonner WM, Martin RF, et al. γ–H2AX foci as a measure of DNA damage: a computational approach to automatic analysis. Mutat Res 2011;711:49–60.