Statistical models for radiation biodosimetry — Poisson or not Poisson?

Jochen Einbeck
jochen.einbeck@durham.ac.uk

in collaboration with
Maria Oliveira (Sant. de Compostela)
Liz Ainsbury and Kai Rothkamm (PHE)
Manuel Higueras and Pere Puig (UAB)
Paul Wilson (Wolverhampton)

Barcelona, 26th October 2015

Durham University

NHS
National Institute for Health Research

Radiation biodosimetry

- Radiation accident or incident
- Triage of individuals requires rapid and reliable procedures to determine the radiation dose
- Biomarkers estimate the dose through radiation–induced changes within cells of the human body

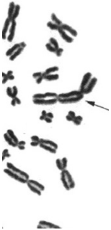
Radiation biodosimetry

- Radiation accident or incident
- Triage of individuals requires rapid and reliable procedures to determine the radiation dose
- Biomarkers estimate the dose through radiation–induced changes within cells of the human body
- Potential biomarkers include
  1. Chromosome aberrations in blood lymphocytes (dicentric chromosomes, micronuclei)
  2. Protein phosphorylation (γ-H2AX)
  3. Gene expressions (microarray or RNASeq)

Cytogenetic biomarkers

- Example: Frequencies of dicentrics (= aberrant chromosome having two centromeres) in $n = 4400$ lymphocytes after *in vitro* whole body exposure with 200 kV X-rays (low LET, sparsely ionising radiation).

	y_{ij}								
x_i	0	1	2	3	4	5	6	7	n_i
1	1715	268	15	2	0	0	0	0	2000
2	638	298	56	8	0	0	0	0	1000
3	247	225	85	37	6	0	0	0	600
4	99	129	92	52	21	5	2	0	400
5	48	88	97	99	36	25	5	2	400



- x_i: dose (in Gy) used to irradiate blood sample i, $i = 1, \ldots 5$.
- y_{ij}: counts of dicentric aberrations in j-th cell of blood sample i, $j = 1, \ldots n_i$.

Dose–response model

- Interested in a model of responses y_{ij} given x_i of type

$$\lambda_i = E(y_{ij}) = h(\beta_0 + \beta_1 x_i + ...)$$

- Count data, so most natural choice is Poisson distribution.

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

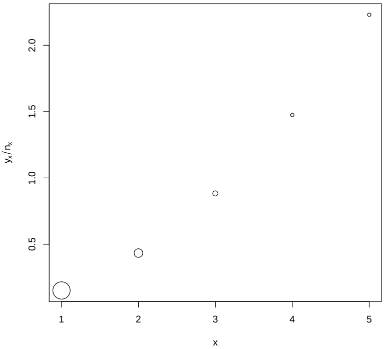
- One can conveniently work at the *aggregated data* level.

Aggregated data

- Let $y_i = \sum_j y_{ij}$. Then the aggregated data are

x_i	n_i	y_i
1.0	2000	304
2.0	1000	434
3.0	600	530
4.0	400	590
5.0	400	892

- Graphically, with circle size $\propto n_i$.

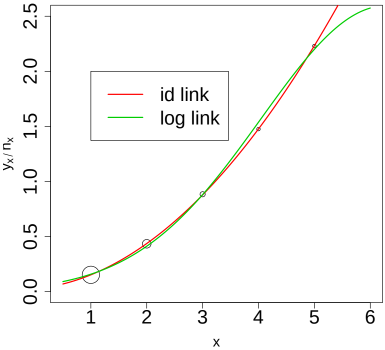


Aggregated data

- Let $y_i = \sum_j y_{ij}$. Then the aggregated data are

x_i	n_i	y_i
1.0	2000	304
2.0	1000	434
3.0	600	530
4.0	400	590
5.0	400	892

- Graphically, with circle size $\propto n_i$.
- Poisson fit, quadratic in dose:

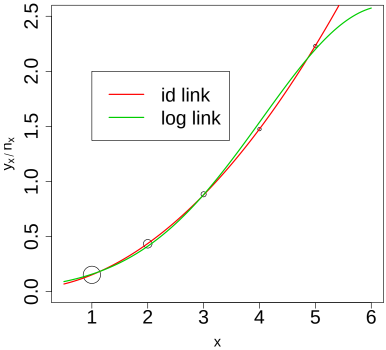


Aggregated data

- Let $y_i = \sum_j y_{ij}$. Then the aggregated data are

x_i	n_i	y_i
1.0	2000	304
2.0	1000	434
3.0	600	530
4.0	400	590
5.0	400	892

- Graphically, with circle size $\propto n_i$.
- Poisson fit, quadratic in dose:

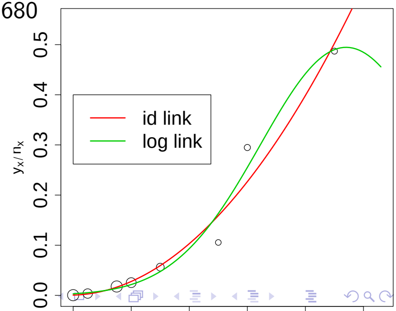


- Brilliant fit of the quadratic id link Poisson model...
- But...

Example 2

- Frequency of dicentrics after *in vitro* whole body exposure to Co-60 gamma rays (low LET)

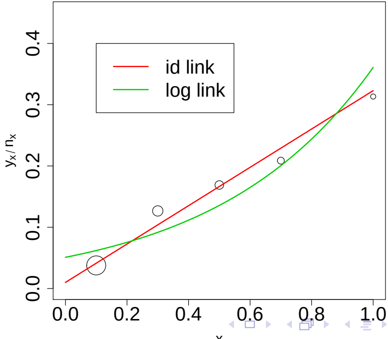
	y_{ij}							
x_i	0	1	2	3	4	5	n_i	y_i
0.00	2591	1	0	0	0	0	2592	1
0.25	2185	8	0	0	0	0	2193	8
0.75	2550	44	1	0	0	0	2595	46
1.00	2231	54	2	0	0	0	2287	58
1.50	1712	96	3	0	0	0	1811	102
2.50	1196	123	7	1	0	0	1327	140
3.00	1070	320	41	6	1	0	1438	424
4.50	895	360	110	25	5	1	1396	680



Example 3

- Frequency of dicentrics after *in vitro* partial body irradiation (50%) with 2.1MeV neutrons (high LET)

	y_{ij}							
x_i	0	1	2	3	4	5	n_i	y_i
0.1	2130	59	9	2	0	0	2200	83
0.3	1088	84	19	6	3	0	1200	152
0.5	875	88	30	7	0	0	1000	169
0.7	679	88	23	8	1	1	800	167
1.0	480	75	27	13	5	0	600	188



Modelling decisions for dose–response curves

1. Predictor: linear or quadratic in dose?
   - linear for high LET, quadratic for low LET
2. Link: Identity or log–link?
   - Identity link preferred by cytogenists as log–link biologically implausible. (The log–link will be preferred by the Statistician, though!).
3. Poisson or not Poisson?
   - Initial graphical evidence seems to indicate that Poisson does not fit well, but how to quantify?
     - Ad–hoc dispersion estimate, for instance, for Example 2 with id–link: $\hat{\phi} \approx \text{Dev}/\text{df}_{\text{res}} = \frac{56.22}{5} = 11.24 \gg 1$;
     - *u*–test;
     - Model fitting: Likelihood, AIC, BIC,...
     - Statistical tests for overdispersion/zero–inflation.

Alternative models

- Several alternative models have been suggested in the literature...
    - Negative Binomial models or Neyman-A — particularly for densily ionising radiation (high–LET)
    - Hermite models — natural justification based on Poisson process
    - Poisson–Inverse Gaussian models
    - Polya–Aeppli models
- A bit neglected (apart from an ad–hoc approach by Dolphin, 1969):
    - Zero–inflated models — biologically plausible especially for partial body irradiation.

Zero–inflation

- ...is a plausible source of overdispersion: Either a cell did not get irradiated (then 0 dicentrics), or it did (then Poisson dicentrics).
- Zero–inflated regression model

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

  where $0 \leq p_i \leq 1$ and $\lambda_i > 0$.
- We use (for now) $p_i \equiv p$ and $\lambda_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$.

Zero–inflation

- ...is a plausible source of overdispersion: Either a cell did not get irradiated (then 0 dicentrics), or it did (then Poisson dicentrics).
- Zero–inflated regression model

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

  where $0 \leq p_i \leq 1$ and $\lambda_i > 0$.
- We use (for now) $p_i \equiv p$ and $\lambda_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$.
- Likelihood

$$\begin{aligned} L &= \prod_{i=1}^{n} \left(1_{y_i=0}(p + (1-p)e^{-\lambda_i}) + 1_{y_i \neq 0}(1-p)e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!} \right) \\ &= \prod_{i=1}^{n} (1-p) \left(1_{y_i=0}(r + e^{-\lambda_i}) + 1_{y_i \neq 0}e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!} \right) \end{aligned}$$

  - cannot use aggregated data
  - no analytic solution

Zero–inflation

- Use M. Oliveira's function `fitcountdist` to fit these models.
- For instance, for data set from Example 2:

```
> Ex2Poi<-fitcountdist(dic~dosevec+dosevec2, data=datavec,
     dist="Poisson", link="identity", start=mustart)
Maximum Likelihood estimation
Nelder-Mead maximisation, 58 iterations
Log-Likelihood: -3748.586 (3 free parameter(s))
Estimate(s): 0.0004975078 0.003037069 0.02412309
AIC= 7503.172     BIC= 7526.144

> Ex2ZIP<-fitcountdist(dic~dosevec+dosevec2|1, data=datavec,
     dist="ZIP", link="identity", start=c(mustart,start0a))
Maximum Likelihood estimation
BFGS maximisation, 166 iterations
Log-Likelihood: -3739.791 (4 free parameter(s))
Estimate(s): 0.0004987518 0.002995455 0.02414752 -1.317389
AIC= 7487.582     BIC= 7518.212
```

Model fitting

- Summary for three example data sets, with log–likelihood $\ell = \log L$ and BIC=$-2\ell + k \log n$. Here k is the number of regression parameters plus additional model paramaters, the latter being given in brackets below.

	Example 1		Example 2		Example 3	
	ℓ	BIC	ℓ	BIC	ℓ	BIC
Poisson (0)	-3806.9	7638.9	-3748.6	7526.1	-2302.1	4621.5
NB (1)	-3806.9	7647.3	-3739.2	7517.1	-2148.7	4323.3
Neyman A (1)	-3806.9	7647.2	-3743.0	7647.2	-2147.0	4319.9
ZIP (1)	-3806.4	7646.4	-3739.8	7518.2	-2155.2	4336.3
Hermite$_2$ (1)	-3806.9	7647.3	-3743.1	7524.8	-2164.8	4355.6
ZINB (2)	-3806.4	7654.8	-3739.1	7526.6	-2143.5	4321.6
Hermite$_3$ (2)	-3808.4	7658.7	-3742.5	7533.3	-2146.5	4327.8
LET	low		low		high	
exposure	whole		whole		partial	

Model fitting with log–link?

- Summary for three example data sets, with log–likelihood $\ell = \log L$ and BIC=-2$\ell + k \log n$. Here k is the number of regression parameters plus additional model paramaters, the latter being given in brackets below.

	Example 1		Example 2		Example 3	
	ℓ	BIC	ℓ	BIC	ℓ	BIC
Poisson (id)	-3806.9	7638.9	-3748.6	7526.1	-2302.1	4621.5
Poisson (log)	-3808.3	7641.7	-3749.4	7527.7	-2323.3	4663.9
ZIP (id)	-3806.4	7646.4	-3739.8	7518.2	-2155.2	4336.3
ZIP (log)	-3807.8	7649.2	-3741.2	7521.0	-2173.3	4372.5
ZINB (id)	-3806.4	7654.8	-3739.1	7526.6	-2143.5	4321.6
ZINB (log)	-3807.8	7657.1	-3740.5	7529.3	-2158.8	4352.2
LET	low		low		high	
exposure	whole		whole		partial	

Model fitting with log–link?

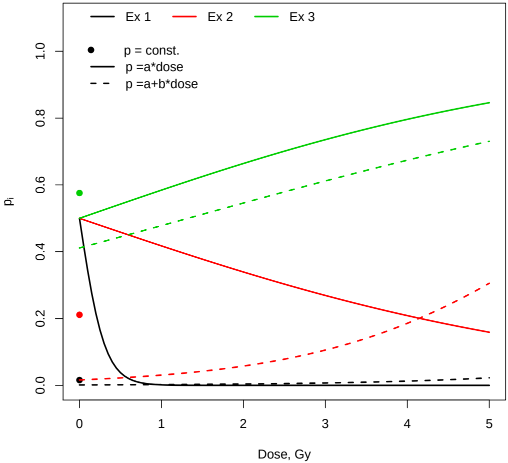
- Summary for three example data sets, with log–likelihood $\ell = \log L$ and BIC=-2$\ell + k \log n$. Here k is the number of regression parameters plus additional model paramaters, the latter being given in brackets below.

	Example 1		Example 2		Example 3	
	ℓ	BIC	ℓ	BIC	ℓ	BIC
Poisson (id)	-3806.9	7638.9	-3748.6	7526.1	-2302.1	4621.5
Poisson (log)	-3808.3	7641.7	-3749.4	7527.7	-2323.3	4663.9
ZIP (id)	-3806.4	7646.4	-3739.8	7518.2	-2155.2	4336.3
ZIP (log)	-3807.8	7649.2	-3741.2	7521.0	-2173.3	4372.5
ZINB (id)	-3806.4	7654.8	-3739.1	7526.6	-2143.5	4321.6
ZINB (log)	-3807.8	7657.1	-3740.5	7529.3	-2158.8	4352.2
LET	low		low		high	
exposure	whole		whole		partial	

- id–link performs generally (a bit) better.

Modelling the zero–inflation parameter

- Zero–inflation parameter can be modelled as a function of dose.



- Models with linear ZIP parameters are biologically plausible and generally lead to a further decrease in BIC.

Score tests for zero–inflation

- $H_0 = Po(\lambda_i)$, $H_1 = ZIP(p, \lambda_i)$ (in other words, $H_0 : p = 0$)
- Score test statistic

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

  where $S(0, \hat{\beta})$ and $J(0, \hat{\beta})$ are the score vector and Fisher information evaluated at $p = 0$ and the Poisson MLE $\hat{\beta}$ for the regression coefficients.
- developed in van den Broek (1995) for the log–link
- adapted in Oliveira et al. (2015) for the identity–link
- Results:

	Example 1	Example 2	Example 3
T (id)	0.92	18.17	387.91
T (log)	1.00	16.89	398.38

- ... to be compared with $\chi^2(1)_{0.95} = 3.84$.

Model choice

- Extensive study: 11 data sets under different exposure scenarios
- Score test results

	exposure	whole					partial					
link	test	low LET			high LET		low LET			high LET		
id	P/ZIP	0.9	18.2	383.6	87.7	61.3	2007.4	1418.3	776.6	416.2	387.9	168.1
	P/ZIP	1.0	16.9	378.7	87.2	47.2	1996.3	1418.0	745.8	421.5	398.4	168.7
log	P/NB	0.9	20.8	1699.9	159.3	136.9	6009.4	3281.0	1210.3	770.6	693.8	285.6
	ZIP/ZINB		1.5	1043.9	47.2	65.0	0.2	1.7	< 0.1	11.5	35.9	36.2

- Recommended model choices for dicentrics

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

Model choice

- Extensive study: 11 data sets under different exposure scenarios
- Score test results

	exposure	whole					partial					
link	test	low LET			high LET		low LET			high LET		
id	P/ZIP	0.9	18.2	383.6	87.7	61.3	2007.4	1418.3	776.6	416.2	387.9	168.1
	P/ZIP	1.0	16.9	378.7	87.2	47.2	1996.3	1418.0	745.8	421.5	398.4	168.7
log	P/NB	0.9	20.8	1699.9	159.3	136.9	6009.4	3281.0	1210.3	770.6	693.8	285.6
	ZIP/ZINB		1.5	1043.9	47.2	65.0	0.2	1.7	< 0.1	11.5	35.9	36.2

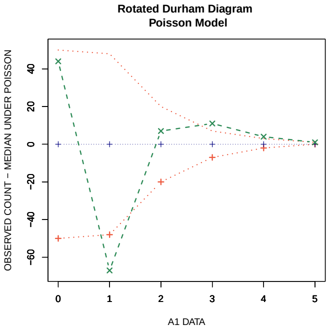
- Recommended model choices for dicentrics

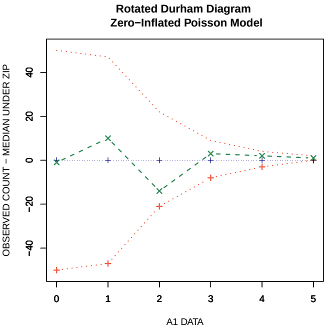
exposure		whole body	partial
LET	low	Poisson/NB	ZIP
	high	NB/Neyman A	ZINB

- For micronuclei, always use ZINB distribution.

Alternative test idea

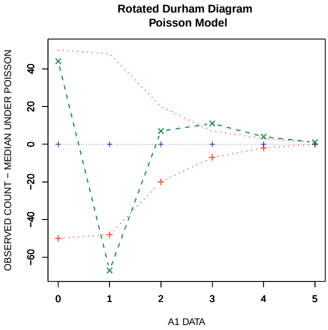
- Plausibility bands on the number of counts under the null hypothesis,
- effectively testing for ‘number–inflation/deflation’,
- confidence limits based on Poisson–Binomial distribution.
- For the data from Example 2,

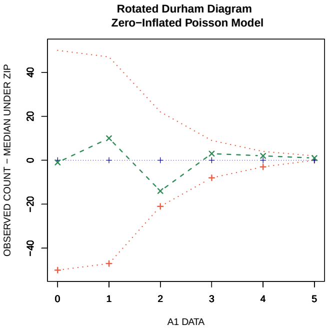




Alternative test idea

- Plausibility bands on the number of counts under the null hypothesis,
- effectively testing for ‘number–inflation/deflation’,
- confidence limits based on Poisson–Binomial distribution.
- For the data from Example 2,

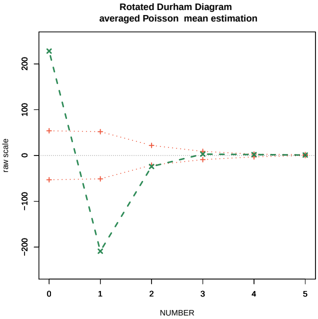


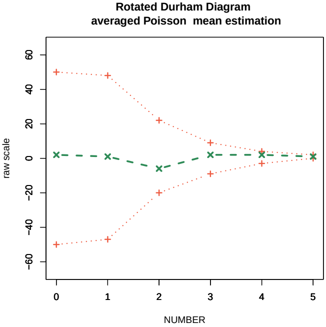


- Zero–inflation often implies 1–deflation...

Alternative test idea

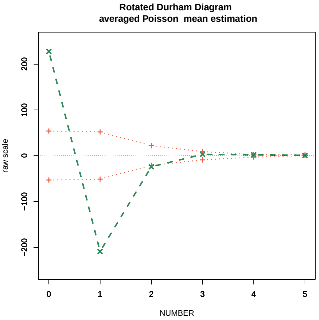
- For 25% partial body Low–LET exposure,

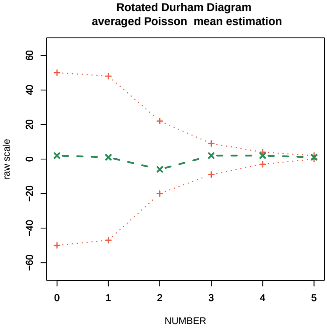




Alternative test idea

- For 25% partial body Low–LET exposure,





- ‘Christmas Eve Test’ (Wilson and Einbeck, 2015).

Random effects?

- ...can be used to capture overdispersion.
- Dicentric counts y_{ij} could be considered to have hierarchical structure, with the 'upper level' i (blood sample) corresponding one–to–one to fixed dose values x_i, $i = 1, \dots, d$, and the lower level j corresponding to cells within samples.
- Hence, random effects on the upper level j induce correlation within subsamples.
- Problem when using additive random effects *and* the identity link: a model of type

  $$\lambda_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + z_i$$

  with random effect z_i, may give $\lambda_i < 0$ for some doses x_i.
  - Biologically and statistically meaningless.
  - Would require complicated constraints....
- Under log–link, less of such problems!

Results for random effects

	Example 1		Example 2		Example 3	
	ℓ	BIC	ℓ	BIC	ℓ	BIC
Poisson (id)	-3806.9	7638.9	-3748.6	7526.1	-2302.1	4621.5
Poisson (log)	-3808.3	7641.7	-3749.4	7527.7	-2323.3	4663.9
ZIP (id)	-3806.4	7646.4	-3739.8	7518.2	-2155.2	4336.3
ZIP (log)	-3807.8	7649.2	-3741.2	7521.0	-2173.3	4372.5
ZINB (id)	-3806.4	7654.8	-3739.1	7526.6	-2143.5	4321.6
ZINB (log)	-3807.8	7657.1	-3740.5	7529.3	-2158.8	4352.2
Pois–RE (id)	-3806.9	7647.3	-3740.7	7519.9	-2301.6	4629.3
Pois–RE (log)	-3808.3	7650.1	-3733.9	7506.4	-2306.8	4639.5
NB–RE (log)	-3808.3	7658.5	-3725.9	7500.0	-2156.5	4347.6

- The log–link leads to less computational issues, and allows for a wider range of models.

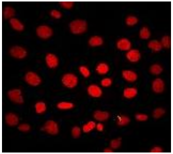
Results for random effects

	Example 1		Example 2		Example 3	
	ℓ	BIC	ℓ	BIC	ℓ	BIC
Poisson (id)	-3806.9	7638.9	-3748.6	7526.1	-2302.1	4621.5
Poisson (log)	-3808.3	7641.7	-3749.4	7527.7	-2323.3	4663.9
ZIP (id)	-3806.4	7646.4	-3739.8	7518.2	-2155.2	4336.3
ZIP (log)	-3807.8	7649.2	-3741.2	7521.0	-2173.3	4372.5
ZINB (id)	-3806.4	7654.8	-3739.1	7526.6	-2143.5	4321.6
ZINB (log)	-3807.8	7657.1	-3740.5	7529.3	-2158.8	4352.2
Pois–RE (id)	-3806.9	7647.3	-3740.7	7519.9	-2301.6	4629.3
Pois–RE (log)	-3808.3	7650.1	-3733.9	7506.4	-2306.8	4639.5
NB–RE (log)	-3808.3	7658.5	-3725.9	7500.0	-2156.5	4347.6

- The log–link leads to less computational issues, and allows for a wider range of models.
- Upper level could be added for data from multiple individuals
  - ...but dicentrics show little inter-individual variation
- Conclusion: Though conceptually attractive, random effect models suffer from practical issues and do not unfold their full power for dicentric data. For other biomarkers, very useful!

γ– H2AX data

- Relatively new technology: Protein biomarker
- Double strand breaks (DSBs) lead to ‘phosphorylation’ of the H2AX protein, yielding γ–H2AX foci.
- γ–H2AX foci are counted using flow cytometers.



- Gives much quicker results than cytogenetic biomarkers, but only if measurement taken with 24 hours of exposure.
- Data from PHE:
  - Blood form several donors (‘multi-individual’)
  - Dose can only be used as a factor, since data collected at 24hr/4Gy and 30min/0.5Gy.
- Two–level hierarchical random effects model (Poisson, log–link, normal or nonparametric random effect) fits and well indicates strong dose effects.

Microarray–based biomarkers

- Currently no established technique which would allow fast ($< 24h$) dose assessment with samples that have been taken at least $24h$ after the radiation incident.
- ‘Pilot’ data available from PHE: 6 donors, 2 genes, 3 dose levels (as continuous variable)
- Gene expression (response) is modelled through Gamma distribution.
- Two–level variance component models with nonparametric random effect fits well: very strong quadratic dose effects, nicely identifiable random effect.
- Gene expressions could be modelled as multivariate response, reducing standard errors and uncertainties...

Conclusion

- For cytogenetic biomarkers, we have found that
  - the Poisson model is mostly inadequate and needs to be replaced by zero–inflated and/or overdispersed models (Oliveira et al, 2015);
  - overdispersion (mainly due to high LET radiation) and zero–inflation (mainly due to partial body exposure) are separately identifiable;
  - in doubt, the ZIP model will do a good job.
- Given a well fitting model, dose can be estimated in a semi-Bayesian inverse regression approach (developed at UAB, Higueras at al, 2015).
- For protein biomarkers, Poisson GLMs with random effects appear useful to to describe a dose *effect* (though current data do not allow to draw dose–response–*curves*).
- For gene expression–based biomarkers, highly promising results using Gamma—GLMs with random effects.

References

(1) Oliveira, M. et al. (2015). Zero–inflated regression models for radiation–induced chromosome aberration data: A comparative study. *Biometrical Journal*, doi 10.1002/bimj.201400233

(2) Higueras, M. et al. (2015). A new inverse regression model applied to radiation biodosimetry. *Proceedings of the Royal Society A*, doi 10.1098/rspa.2014.0588

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

(4) Wilson, P. and Einbeck, J. (2015). A simple and intuitive test for number-inflation or number-deflation. In: Wagner, H. and Friedl, H. (Eds). Proc's of the 30th International Workshop on Statistical Modelling, Linz, Austria, 6-10 July 2015, Vol 2, pp 299–302.

[References to data sources given in (1)]