# Modelling Incident Reporting:
# - tackling poor quality count data

**Chris Mainey**
PhD Student – Primary Care and Population Health - UCL
Intelligence Analyst – University Hospital Birmingham NHS FT

christopher.mainey.14@ucl.ac.uk    chris.mainey@uhb.nhs.uk

**Supervisors:**

- Prof. Nick Freemantle - UCL
- Dr. Milena Falcaro - UCL
- Prof. Simon Ball - UHB

# Summary

- Explain what I'm doing
- Dataset construction
- Initial modelling:
  - Poisson Models
  - Parameterisation
  - Over-dispersion
- Random Effects models
- Generalized Additive Models

# Incident reporting in Healthcare

- **“Incident”** – Event or situation where, or with the capacity to lead to, patients or staff may be harmed

- Reported locally, submitted to National Reporting and Learning System (NRLS)
  - Variety of other systems – confusing landscape

- Philosophical and logistical problems:
  - Definitions
    - What is an incident?
    - Focus of incident: patient, staff, omission, potential for problem…?
  - Fidelity of reporting
    - Under-reporting
    - Missing data

# NRLS quantitative data

- NRLS is primarily qualitative
  - The strongest ‘signal’ is in the free-text descriptions
- Approx. 1.8 million reported per year
- ‘Severe harm’ and ‘Death’ reviewed
  - <1% reports
  - Reporting of other harm levels not mandatory
  - Current national analyses ignore the majority of the dataset
- Unclear outcomes
  - are more incident reports a bad thing?
  - High error rate or good awareness of risk/mature reporting culture?

# Theory

- I proposing that:

$$Incident\ reports = f(Exposure) + f(Culture) + error$$

- Exposure = opportunity for error (e.g. large v.s. small organisations)
- Culture = awareness, reporting behaviour

- Both ‘latent variables’
  - Can’t be directly measured
  - Looking to identify proxy measures for exposure

# Data Loading & management

- Monthly extract from NHSI team, based on date received/ by NRLS
- Received as ‘csv,’ process for formatting and extracting to SQL Server.
- Error checking: nulls values, missing data, merged organisations, duplicates,
- Aggregated and joined with additional dataset

# Additional data set: HES

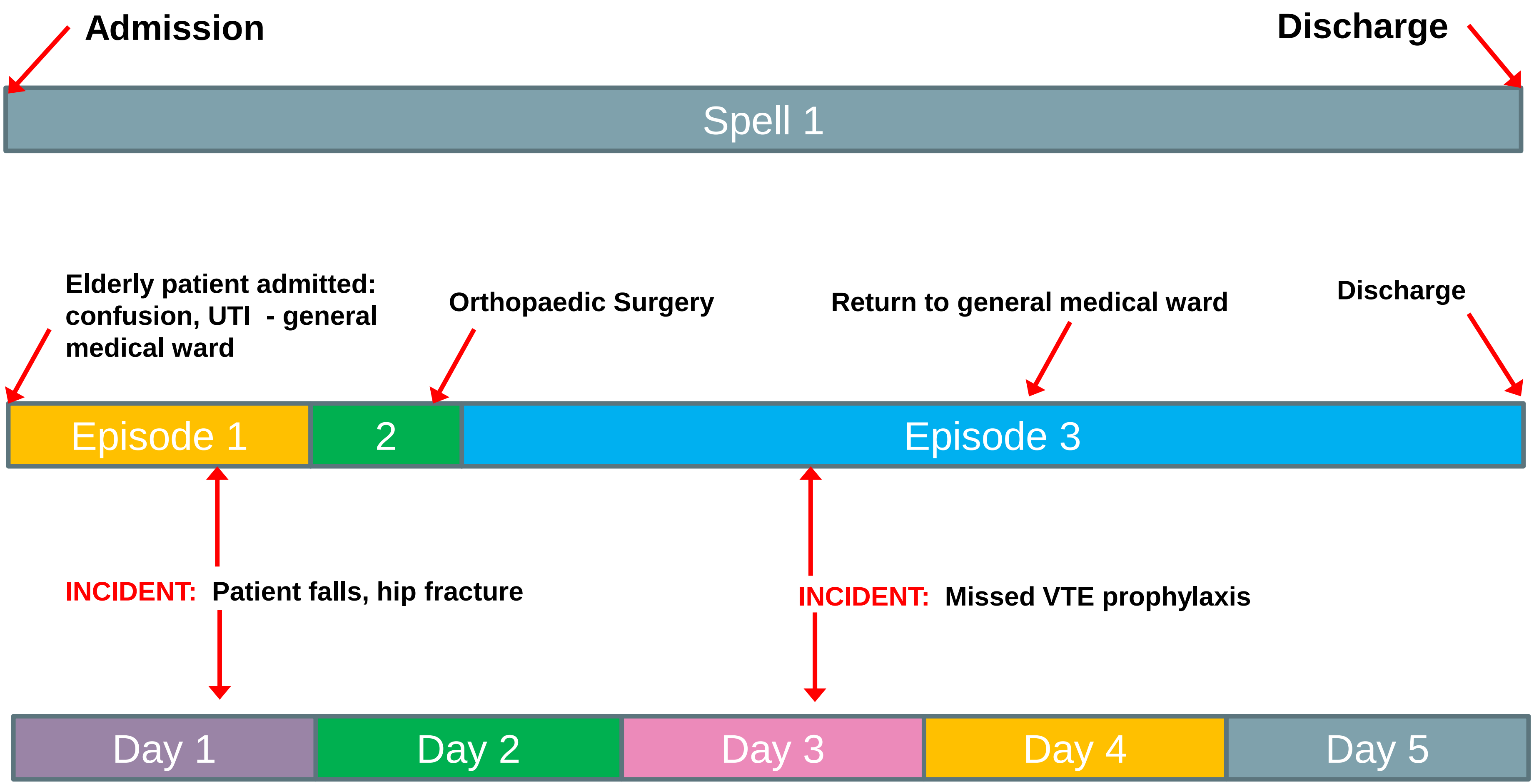
- Lit. review suggest NRLS categorical data not sound modelling

- Hospital Episode Statistics (HES)
  - In-patient & Outpatient records,
  - Demographics and case-mix factors

- Directly linkage not possible:
  - No identifiers
  - Not collected in the same ‘units’
    - patient flow in HES proxy for size/exposure
    - Probabilistic linkage not appropriate
  - IG rules

- Construct count dataset, per organisation, per month
  - Contingency table / “panel” data
  - Counts of Incidents, and ‘bed-days’ in demographic groups

# What's a bed-day?




For this patient, we've counted 5 units of exposure, 2 events occurred

# Modelling approach

- Count data:
  - Properties or count data:
    - Discrete
    - Bounded at zero
    - Likely skewed

- Generalized linear Model framework (Nelder &Wedderburn, 1972):
  - Poisson Regression:
  - $\log(\mu) = \mathbf{X}\beta$
  - $\log(y_i) = \beta_0 + \beta 1_i X1 \ldots \beta p_i Xp$

- Work all conducted in R, using standard 'glm'

# Models (generally)

- $\text{incidents} = \text{Age (IP)} +$
  - $\text{Sex (IP)} +$
  - $\text{Co}-\text{morbidity score (IP)} +$
  - $\text{Adm. Method(IP)} +$
  - $\text{Age (OP)} +$
  - $\text{Fin.Year} +$
  - $\text{Time}-\text{trend}$

- Multiple categories of each parameter
- Incidents during 2011/12 – 2015/16
- Fiscal year as categorical, Time trend as natural cubic-spline

# Parameterisations

1. Proportions:

$$\frac{Bed\text{−}days\ Age\ Group\ z}{Total\ Bed\text{−}days}$$

- All on same scale 0 – 1
- Lose size or effect – bed-days as ‘offset’
- Perfect multi-collinearity / identifiability issues:
  - Several sets of parameters summing to 1:  not estimable.
  - Need to drop one level.

# Parameterisations

2. Count:

$$\log(Bed\text{–}\ days\ \ in\ Age\ Group\ z\ )$$

- Poisson distributed covariates
- Should be log-transformed:
  - Maintain linearity on the scale of link-function
  - More easily estimated by software
- Size element is maintained and does require an ‘offset’
- All parameters can be fitted as no collinearity issue

# Parameterisations

3. Quantiles of covariate distribution:

$$quantile(count\ of\ Bed - days\ in\ Age\ Group\ z)$$

- Values: median, min, 0.05 0.25, 0.5, 0.75, 0.95, max

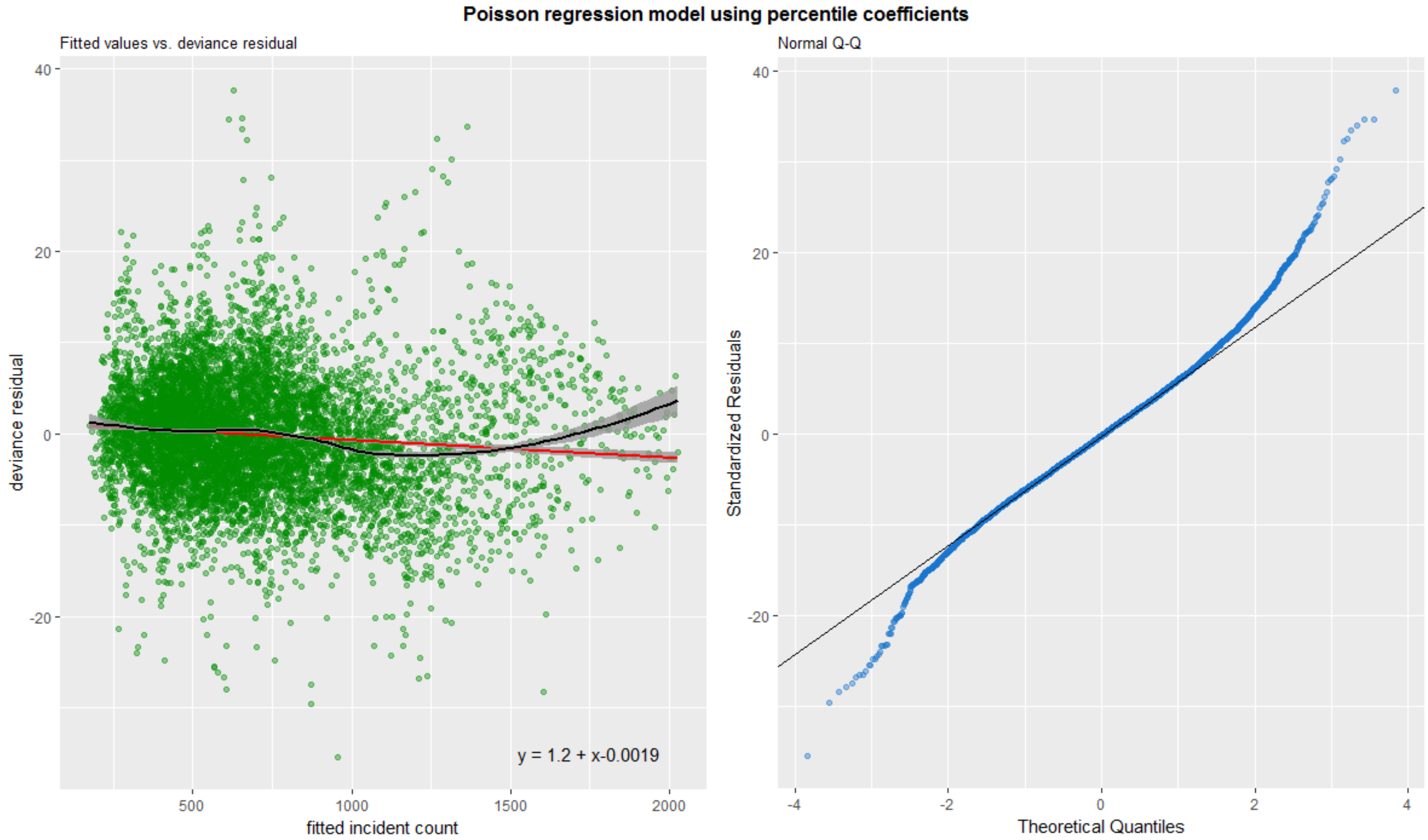
- Per organisation, per month, description of distribution
- Size element is lost and ‘offset’ is required
- All parameters can be fitted as no collinearity issue
- Computational burden: estimated ~340 days in single threaded R session.
  - Reduced to 1.4 days through parallelisation, and efficient loop coding

# Fitted models

- Poor fit for each Poisson model
  - Chi-sq tests on deviance vs. residual degrees of freedom
  - High AIC
  - All parameters ‘significant’ at 95%
  - Heteroscedasticity

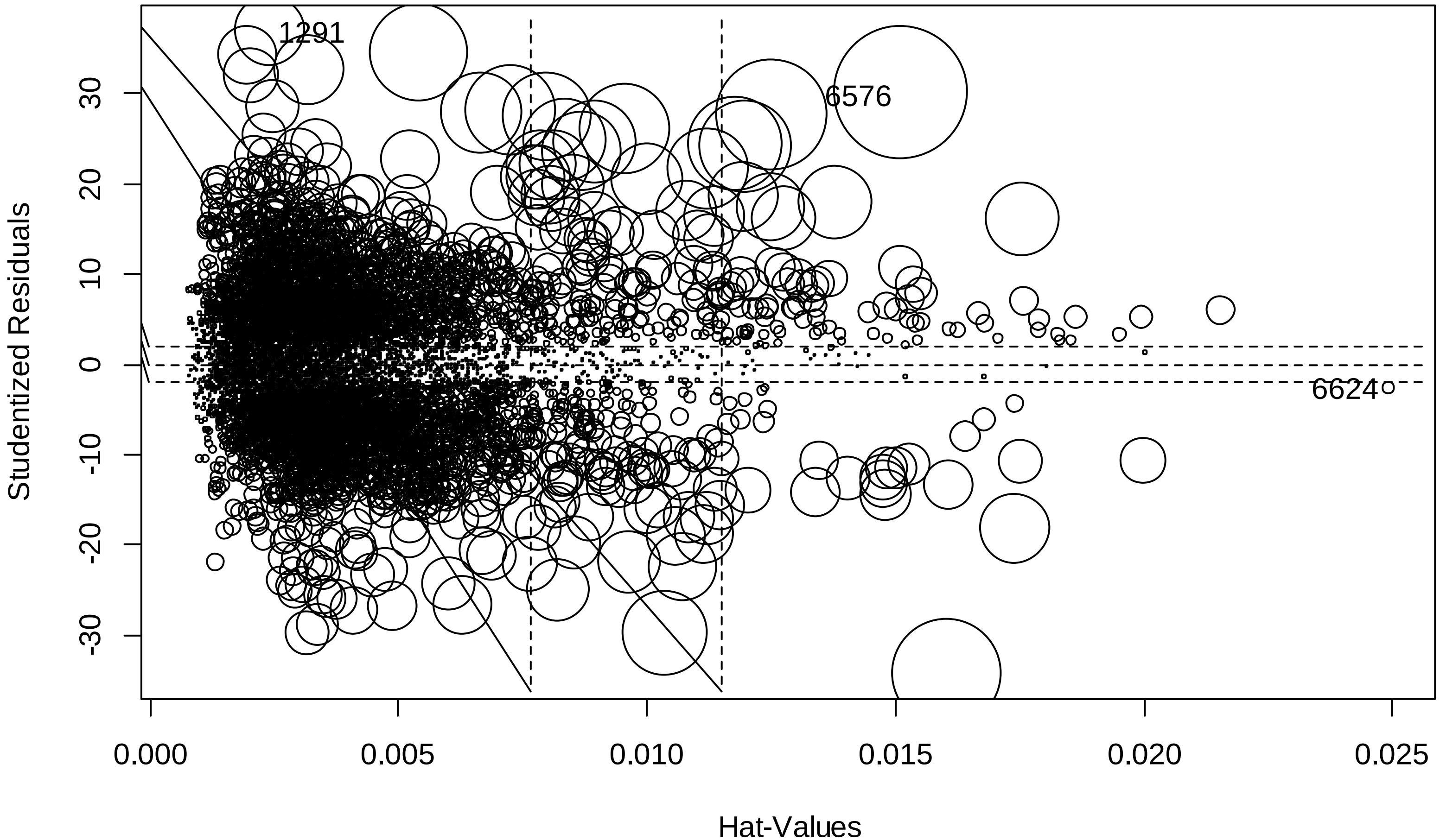
- Over-dispersion:
    - Poorly specified linear part of model
    - Presence of outliers
    - Clustering

# Outlier detection




- Number of rounds of screening using fitted, residual and influence values.
- Non-constant variance
- Data excluded where valid reason, e.g. missing HES data

- Useful R function: ‘influencePlot’ (car package).




- Size: observations proportional to Cook's distances
- Highlights outlying results

# Alternatives

- **Quasi-likelihood model:**
  - Scaled likelihood function, allows over-dispersion adjustment of standard error.

- **Clustering:**
  - Data are sets of 60 repeated measures at clusters (hospitals)
  - Correlation structure required, as within cluster variance is not accounted for.  Poisson GLM assumes independence.
  - **Random-intercept**:  allow intercept to vary for each organisation, acknowledge clustering, but estimating fixed effect for all
  - **Random-intercept & slope:**  allows intercept to vary based on another parameter.  In this case, it fiscal year.

# Alternatives (2)

- Quasi-models
  - Large impacts on error, better estimates of significance
  - No AIC to compare
  - Ignores correlation structure

- Random effects
  - Significant drop in AIC with both models, with random intercept and slope giving lowest.
  - Replicated across all parameterisations

# Questions:

- How would I best test which parameterisation is ‘better’?

- Any other thoughts on model structure?
  - Alternative approaches
  - Random effects structures

# Smoothed models

- The data are ‘noisy’ but show some general trends.
- Variables might be better modelled as smoothed functions

- Artificial divides in covariates e.g. age to allow parameterisation
- What if we could pool covariates into a ‘smoothed surface’ for fitting?

- **Generalized Additive Model (GAM)** (Hastie & Tibshirani, 1990)

# GAMs

- GLM with linear predictor is a sum of smooth functions of covariates of general form:

$$g(\mu_i) = \mathbf{A}_i\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots$$

- Where:
  - $\mu_i \equiv E(Y_i)$, and $Y_i \sim EF(\mu_i, \phi)$
  - $\mathbf{A}_i$ row of model matrix for strictly parametric components and $\boldsymbol{\theta}$ corresponding parameter vector
  - $f_j$ smoothed covariates of $x_k$

- Flexible specification due to smoothers, but now need to:
  - Represent smooth functions in some way
  - Choose how smooth they should be

(Wood, 2017)

# R package: mgcv

- Fits GAMs by penalized MLE
- Variety of smoothers recognised
  - Cubic splines
  - Thin-plate splines & ‘soap film’ smooths
  - Tensor products
  - Random effects as Gaussian Random Fields
- Smoothness estimation through generalized cross-validation
- Estimation of scale parameter, % deviance, AIC
- Best performance so far

# Random Forest

- Ensemble method combining:
  - Regression Trees
  - Bootstrap aggregation (‘bagging’)
- Large number of trees grown and mean predictions used
- Non-linear models
- Feature selection
- Correct for regression trees tendency to over-fit
- Encouraging results.  Comparable/better than GAM

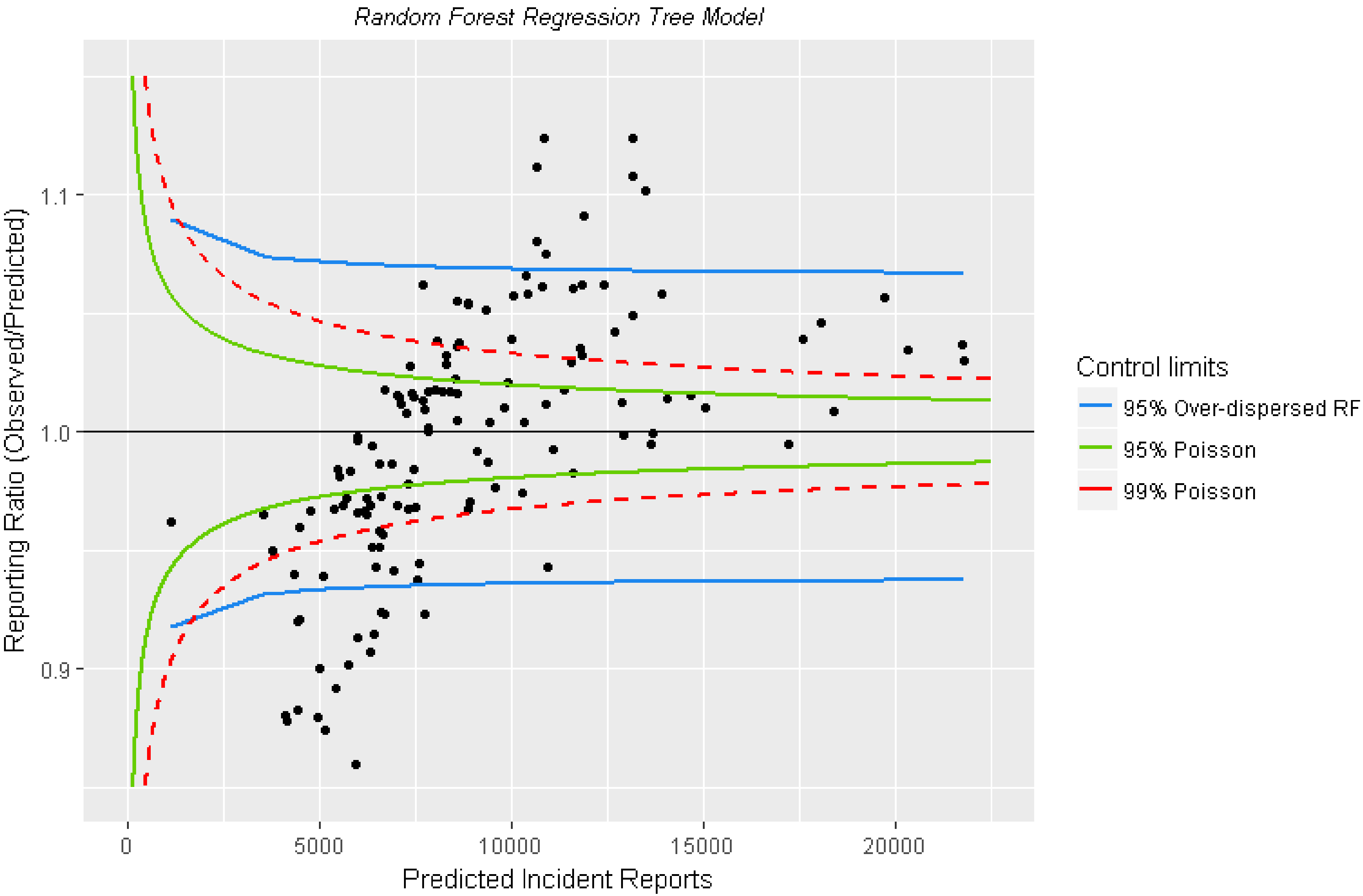
# Presentation

- Presenting and discussing the model coefficients
  - Difficult to understand parameterisations
  - Linear predictors or IRR

- Preferred parameterisation(s)

- Want to show differences between organisations:
  - Too many to fit fixed effects/use parameter estimate
  - Random effects?  Harder to interpret
  - Observed v.s. predicted:
  - Funnel plot common in sector context

# Figure 1: Funnel Plot of Incident reporting ratio




Over-dispersed control limits based on Spiegelhalter et al. (2012)

## Questions:

- How would I best test which parameterisation is ‘better’?

- Any other thoughts on model structure?
  - Alternative approaches
  - Random effects structures

- Any thoughts or objections to GAMs?

- Any experience with or advice about Random Forests?

# Useful References


- Wood, S. (2017) **Generalized Additive Models: An Introduction with R, Second Edition.** CRC press
  - Best general introduction to linear, generalized linear and generalized additive models I've found. Lots of examples in R and recently published.

- Nelder, J.: Wedderburn, R. (1972). "Generalized Linear Models". *Journal of the Royal Statistical Society. Series A (General)*. Blackwell Publishing. **135** (3): 370–384.
- Hastie, T. J.; Tibshirani, R. J. (1990). **Generalized Additive Models**. Chapman & Hall/CRC.
- Breiman, L. (2001). **Random Forests**. Machine Learning. **45** (1): 5–32.
- Spiegelhalter, D. J., C. Sherlaw-Johnson, M. Bardsley, I. Blunt, C. Wood and O. Grigg (2012). **"Statistical methods for healthcare regulation: rating, screening and surveillance."** Journal of the Royal Statistical Society: Series A (Statistics in Society) **175**(1): 1-47.