



Building predictive models with HES data

Readmissions in HED

https://github.com/chrismainey/Readmissions_in_HED_NHSR2020

Chris Mainey

Senior Data Scientist

University Hospitals Birmingham NHS FT

chris.mainey@uhb.nhs.uk

[@chrismainey](https://twitter.com/chrismainey)



1 / 22



Healthcare Evaluation Data (HED)

www.hed.nhs.uk

- Online hospital benchmarking system
 - Statistical models and analysis tools
 - Activity, Mortality, Re-admissions, Length-of-Stay, Market-share etc.
 - Built by Informatics team at University Hospitals Birmingham NHS Foundation Trust
 - Used by ~60 NHS and other organisations
 - Training and support, including R
-
- Using national NHS data, including HES, ONS mortality, central returns, NRLS and others

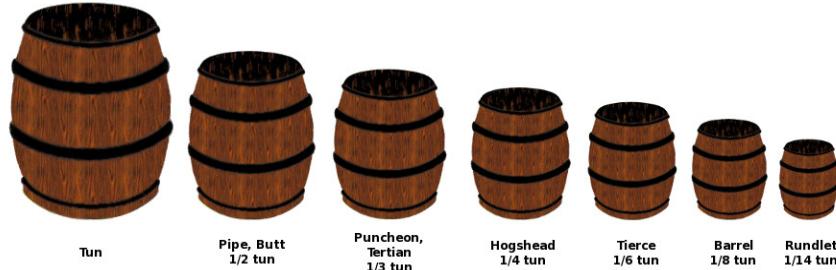


2 / 22



Casemix-adjusted indicators

How can we compare indicators across different centres/units?



By Grolltech; Own work: [CC BY-SA 3.0](#) [Link](#)

- Aggregated patients in different sizes units
- Each patients is different
- Consider biases approach
- Important variables may be:
 - Age profiles
 - Elective / Non-elective balance
 - Seasonality

3 / 22



Indirectly-standardised ratio

- Adjust all to the expected average risks
- Commonly use a regression model to estimate effects of predictors.
- Then use model to predict the risk of event for each patient.
- We can compare our predicted risks to observed events
- **Relative risk ratio:**
$$\frac{\sum \text{events}}{\sum \text{risk}}$$
- Compare our relative risk ratio to the standard (usually 1, or multiples like 100)

4 / 22



Case-study: Relative-Risk Readmission

*Readmission to any acute provider within 30-days of discharge from another.
Indexed to discharge from the first organisation.*

- Major variables relate to age, sex, admission method, diagnosis, comorbid conditions.
- How we parametrise these variables affects quality of model.
 - E.g. Age as continuous? Assumes effects of age are constant.
 - What if it's not? Binning or transformations?
- Regression assumes all points are independent, **this is not true here:**
 - Patients at hospital X more like 'hospital X' patient than 'average patient'
 - Clustering

5 / 22

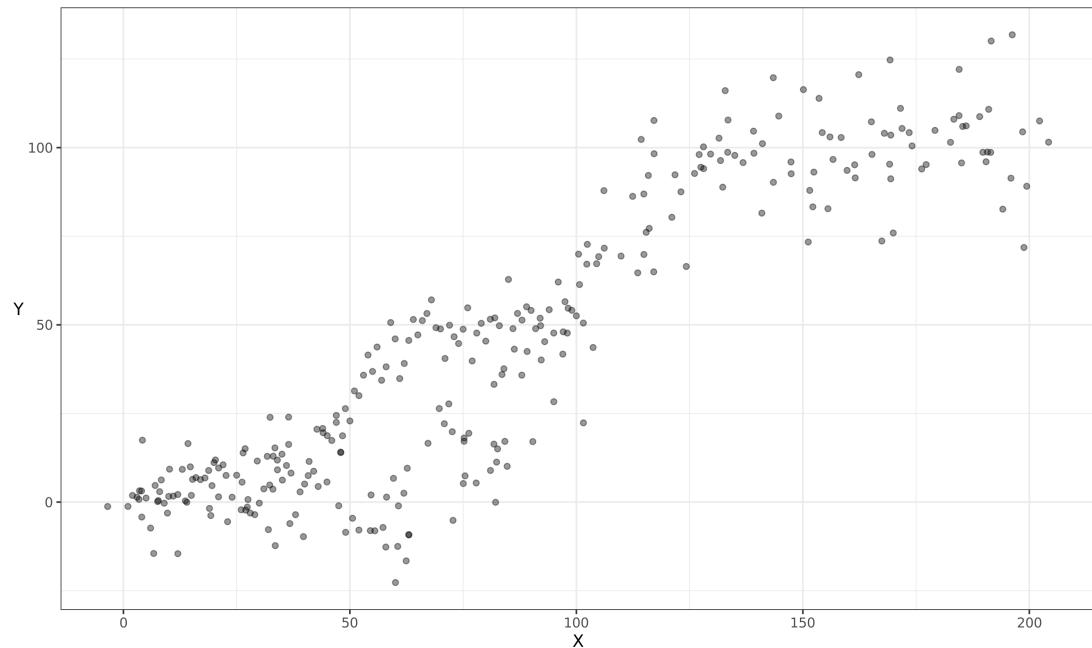


Non-linear data:

What if the relationship between X and Y varies across the range?

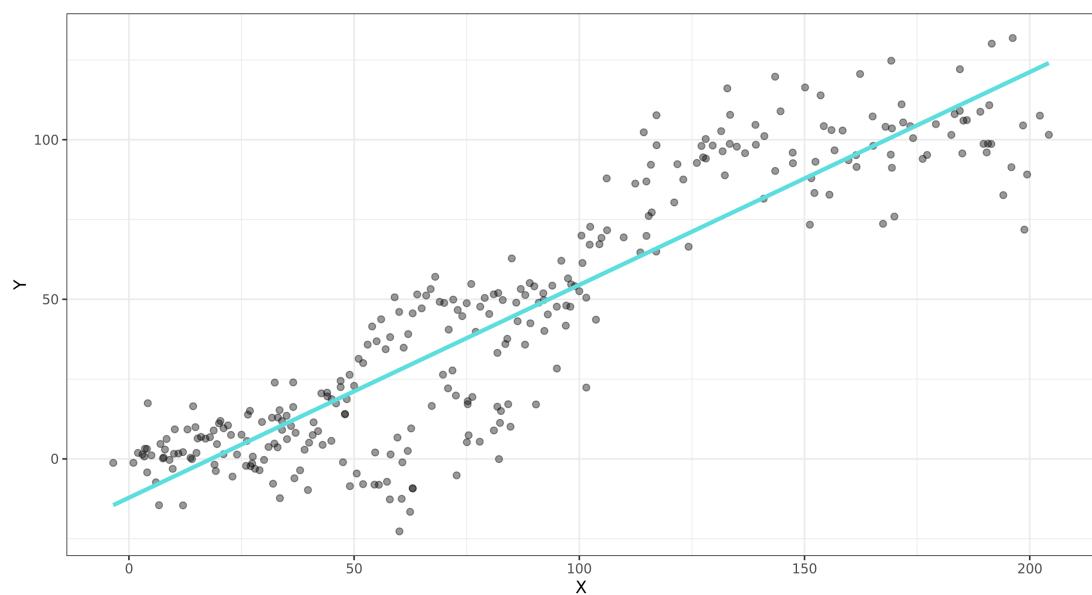
6 / 22

What about nonlinear data? (1)



7 / 22

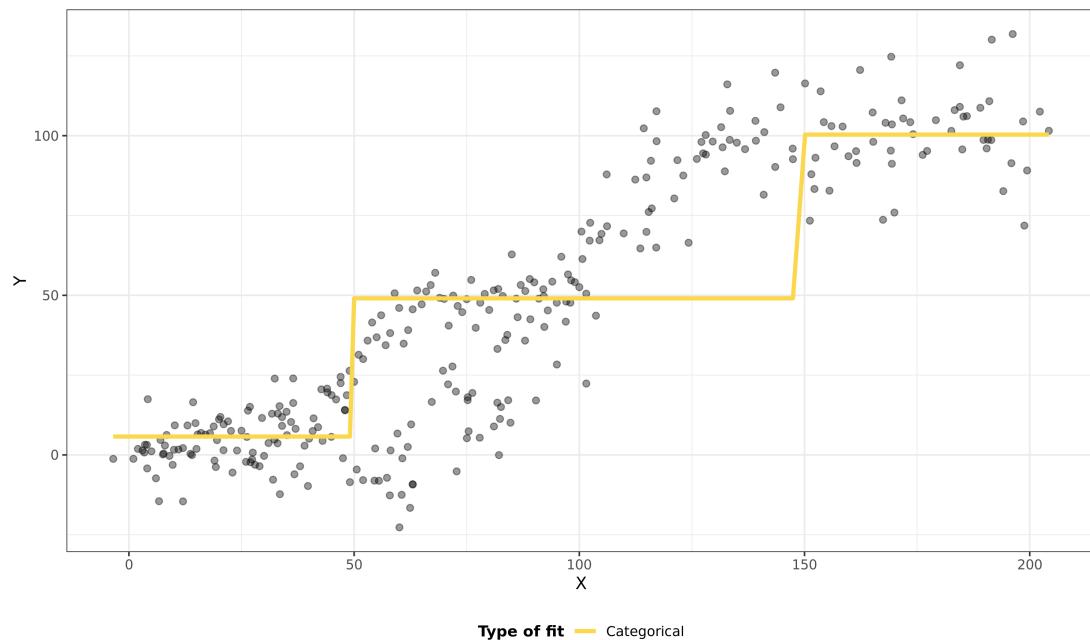
What about nonlinear data? (2)



Type of fit — Linear

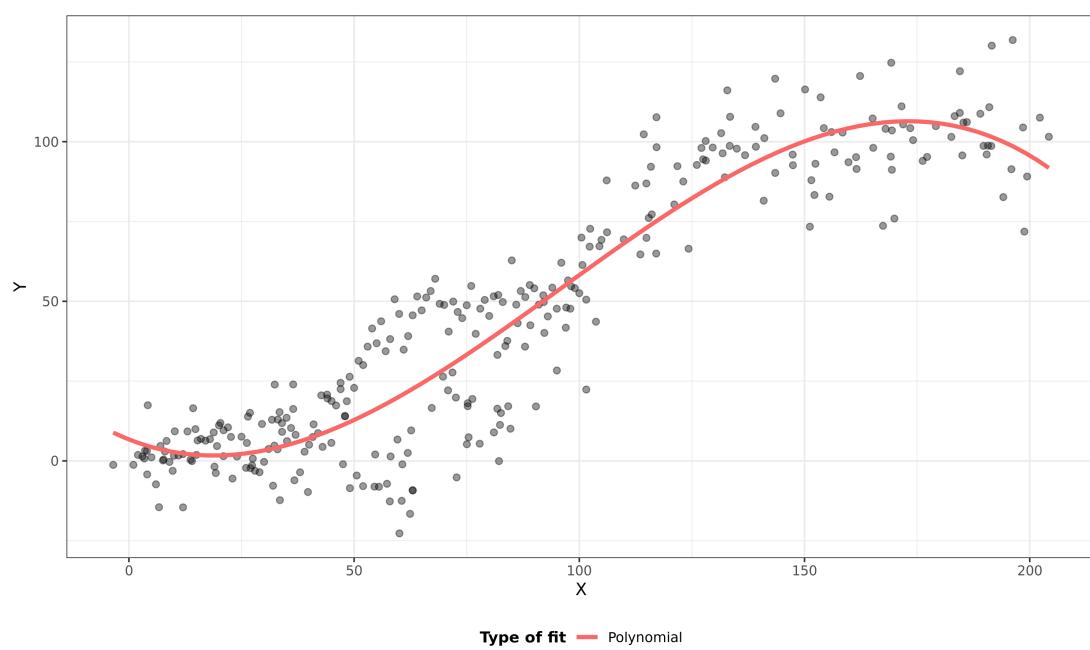
8 / 22

What about nonlinear data? (3)



9 / 22

What about nonlinear data? (4)



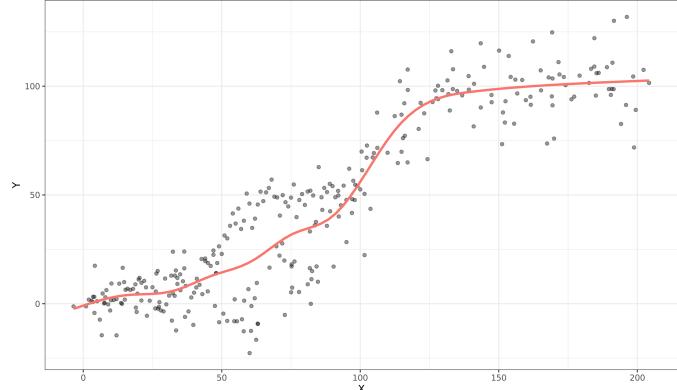
10 / 22



GAMs + Splines

- Smooth, piece-wise polynomials, like a flexible strip for drawing curves.
- Joined at 'Knot points' between each section
- This can then be a Generalised Additive Model
- Essentially: a regression on the sum of smoothers

$$y = \alpha + f(x) + \epsilon$$



11 / 22



GAMs in R

Prof. Simon Wood's package is de-facto standard

```
library(mgcv)
my_gam <- gam(Y ~ s(X, bs="cr"), data=dt)
```

- `s()` control smoothers
- `bs="cr"` telling it to use cubic regression spline ('basis')
- Knots (or equivalent) are set by `k` argument, e.g. `k=10`

12 / 22



Model Output:

```
summary(my_gam)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## Y ~ s(X, bs = "cr")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.9659    0.8305   52.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df F p-value
## s(X) 6.087 7.143 296.3 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.876  Deviance explained = 87.9%
## GCV = 211.94  Scale est. = 206.93 n = 300
```

13 / 22



Clustering

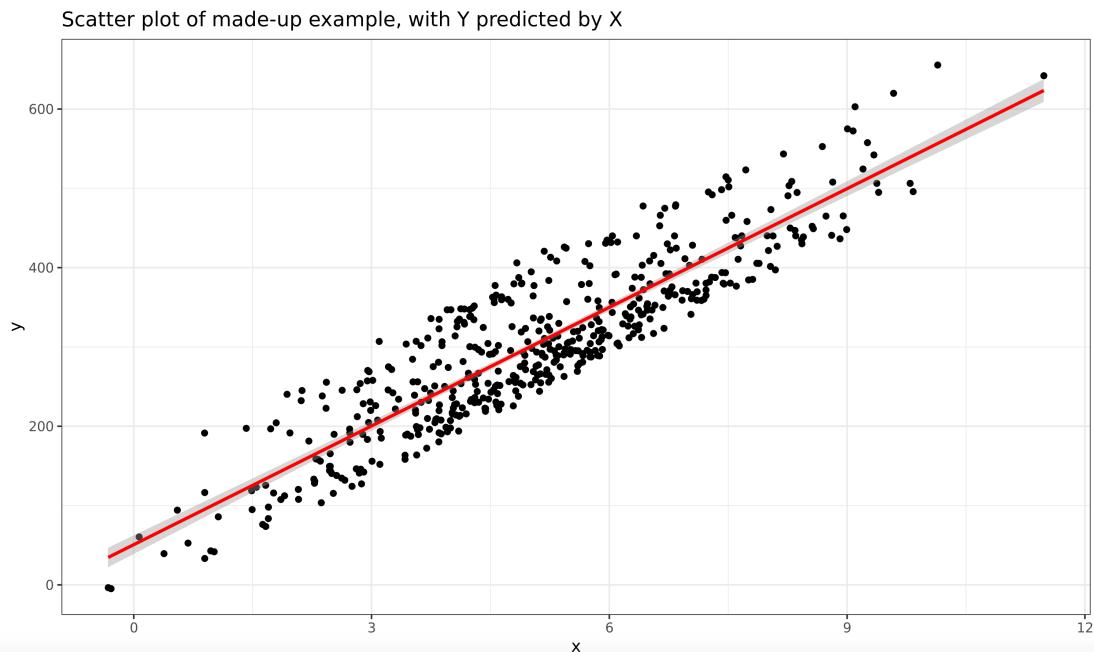
Data collected in unit/centrers, not at random in the population

14 / 22



'Random effects'

Lets imagine we have a big cloud of data points that look like this:

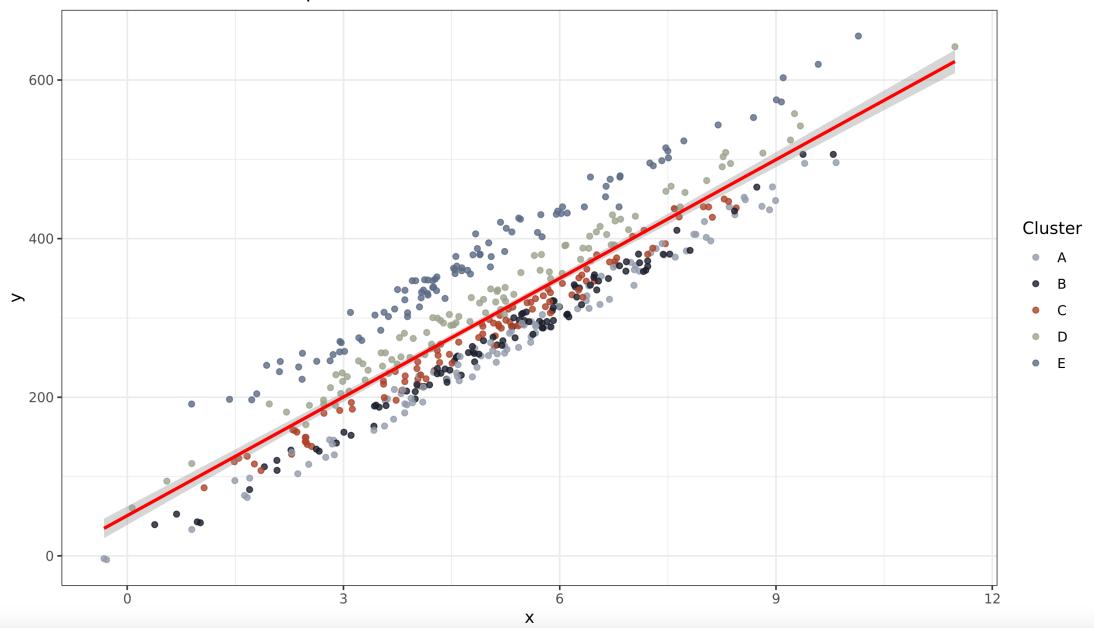


15 / 22

Random effects (2)

If we assume all points are independent, the previous model was fine, but...

...what if the data are repeated measures from clusters

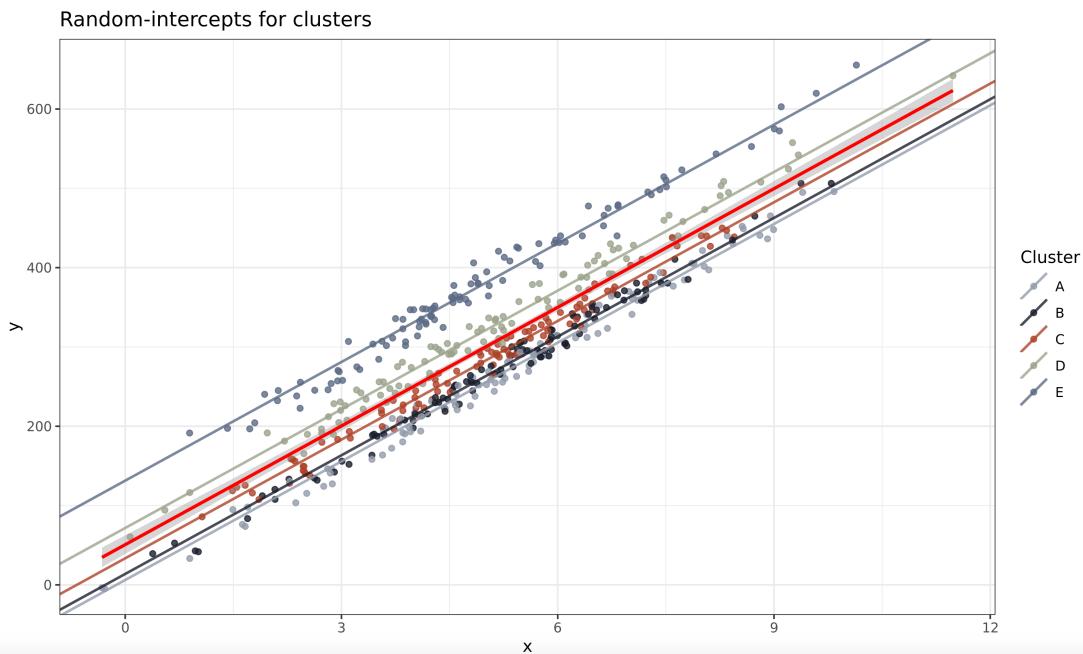


16 / 22



Random effects (3)

If we assume all points are independent, the previous model was fine, but...



17 / 22



Random effects (4)

So we end up with a 'random-intercept' model:

```
library(lme4)
my_ri_model<-lmer(y~x+(1|clust), data=dfc)

summary(my_ri_model)

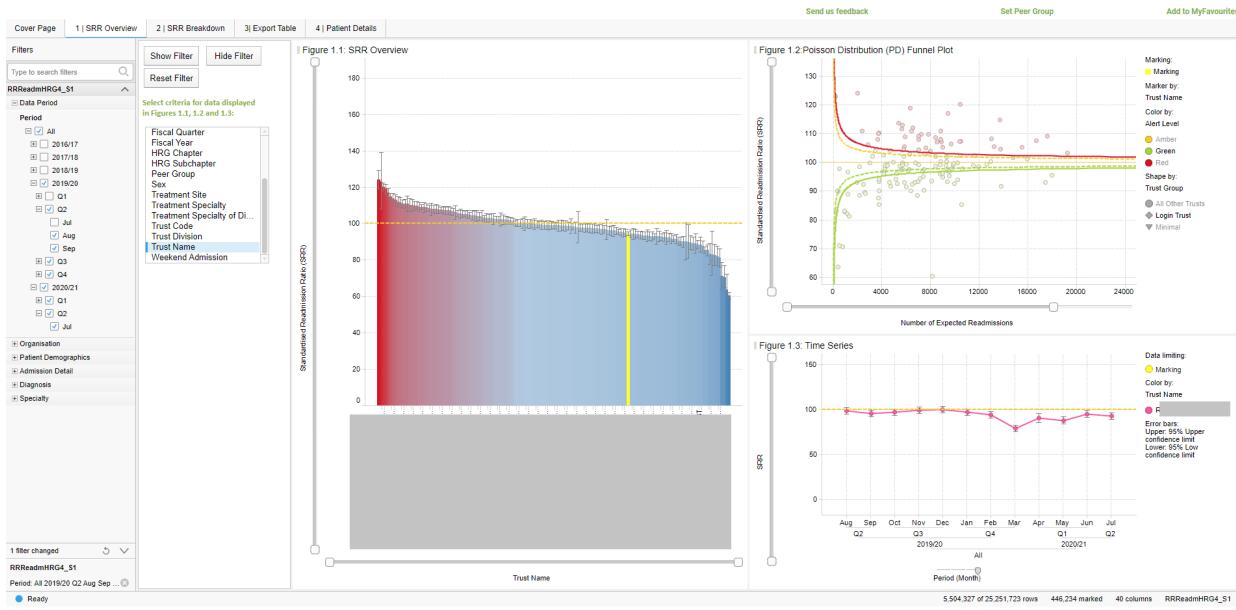
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ x + (1 | clust)
##   Data: dfc
##
## REML criterion at convergence: 3955.8
##
## Scaled residuals:
##    Min     1Q   Median     3Q    Max 
## -2.73388 -0.79825  0.01282  0.83659  2.80617
## 
## Random effects:
##   Groups   Name        Variance Std.Dev.
##   clust    (Intercept) 2651.3   51.49
##   Residual           151.1   12.29
##   Number of obs: 500, groups: clust, 5
## 
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 43.4078   23.0788  1.881
## x          51.3188    0.2839 180.796
## 
## Correlation of Fixed Effects:
##   (Intr) 
##   x      -0.062
```

18 / 22

How do we use it?



Web-based, interactive 'modules' that users can interrogate:



19 / 22

...but HES is pretty big, right?



YES! Yes it is, so required special handling:

- Memory efficiency and speed - `data.table` package
- Only load section required for each model:
 - Use database (SQL Server) for what it's designed for!
 - Stratified by each HRG4 sub-chapter
 - Sparse model matrix
- Parallelisation - `doParallel` - better on Linux, speaking of which:
- Linux! - Set up a VM on server, RStudio Server.
- Optimised functions, like `bam()` in `mgcv`, `bigglm()`

20 / 22



Journey in HED

- HED used SAS for many years to build regression models.
- CM had PhD project funded by UHB that allowed space to learn R
- CM was useless for the first 18-months!
- Then started translating 'broken' SAS models to R
- Initially used CM's (annotated scripts)
- Not sustainable: couldn't pass to other analysts, not fault tolerant, no metadata
- Built R package - MB primarily translated scripts
- R package building encouraged use of Git source control
- Model management database, powered by functions in R package

21 / 22



Summary

- R is a powerful tool for building case-mix adjustment models
- Important to understand your data generation mechanism before modelling
- Regression approach, common in indirect standardisation, have assumptions
- When modelling hospital readmissions for HED:
 - Specific modelling of non-linear relationship increased fit, using `gam()`
 - If clustering affects your data, random-intercepts may be helpful
- When building methods, remember it is marathon, not a sprint
- Use R in its right place in the pipeline
- Efficient handling is essential
- Building R packages and using source control has been great help

22 / 22