SLS-DSU
Scottish Longitudinal Study Development & Support Unit

Using e-DataSHIELD to run joint analysis from three National Statistics Agencies in the UK

Gillian Raab

Administrative Data Research Centre – Scotland

Administrative Data Research Network

An ESRC Data Investment

Introduction to ADRC-S

- Brings together major Scottish centres: Aberdeen, Dundee, Edinburgh, Glasgow Herriot-Watt, St Andrews, NHS working closely with Scottish Government.
- Leading experts in the Law, linkage, computer science (eg Natural Language Processing, Machine Learning) and significant research areas
- Builds on existing services
- Significant programme of public engagement
- Will exploit Scotland's rich admin. data (historical and contemporary)

Summary

- Introduce the three UK Longitudinal Studies (LSs)
- e-DataSHIELD routines
- Practical problems of implementation
- Some results from a project comparing Scotand and Northern Ireland

The UK LSs common features

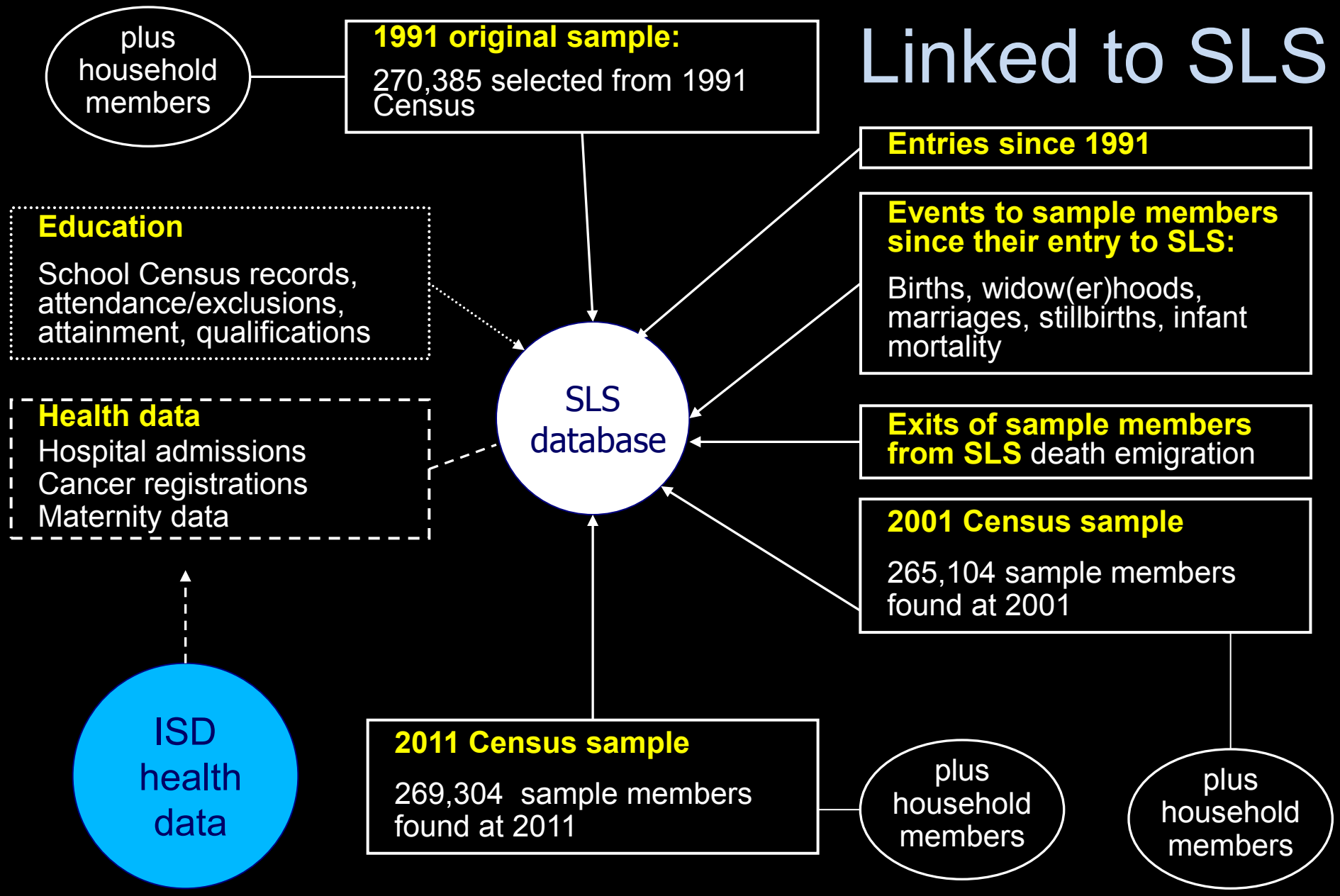
- ONS-LS (England and Wales)
- SLS (Scotland)
- NILS (Northern Ireland)

- All use data from the UK decennial Censuses
- Census forms very similar – minor differences
- All are based on a number of secret birthdays
- Individuals are linked over time at Censuses and from administrative data
- All are held in hyper-secure settings

The three UK LSs differences

- Timing
  - ONS-LS started 1974, 5 Censuses 1971-2011
  - SLS started 1995, 3 Census 1991-2011
  - NILS started 1999, 2 Censuses 2001-2011
- Different numbers of secret birthdays mean different sampling fractions
  - 1% ONS-LS 5%SLS 20% NILS
- The LSs have different linked data sources
- Each Census administered by different National Agency

Linked to SLS



Joint analyses from the three LSs

- What for?
  - To increase the sample available
  - To carry out inter-country comparisons
- Problems with existing infrastructure
  - Analyst will need to travel to safe settings in London, Edinburgh and Belfast to carry out each analysis
  - Only limited data summaries can be taken out of the safe settings

e-DataSHIELD

- After Lyon workshop in 2011 we realised that Datashield could provide an answer
- But National Agencies will not link their studies to a central server- no matter how secure
- Solution: exchange summaries by email.
- e-DataSHIELD (eds) routines written
- Methodology approved by the three LSs.
- Initial project agreed between SLS and NILS

First project

- Comparing mortality by religion for Scotland and Northern Ireland
  - Roman-Catholic vs protestant is main interest
  - Sectarian conflict in NI, less so in Scotland now
  - But this is disputed
- e-DataSHIELD steps (eds routines)
  - Agree data to use
  - Agree models to fit
  - Harmonise data with eds routines
  - Fit models using eds routines
  - analyse results with eds routines

First steps

SLS safe setting ISC1

- Prepare data
- Add a variable for study number=1 to allow interactions
- Export file with summary of structure of data for fitting

NILS safe setting ISC2

- Prepare data
- Add a variable for study number=2 to allow interactions
- Export file with summary of structure of data for fitting

Analysis computer (AC) or an ISC(could be anywhere)

- Read in the two data structures
- Compare and list common variables
- Check that they are of the correct type (e.g. numeric, factor)
- And if factors that levels, labels and contrasts agree

Harmonising the data

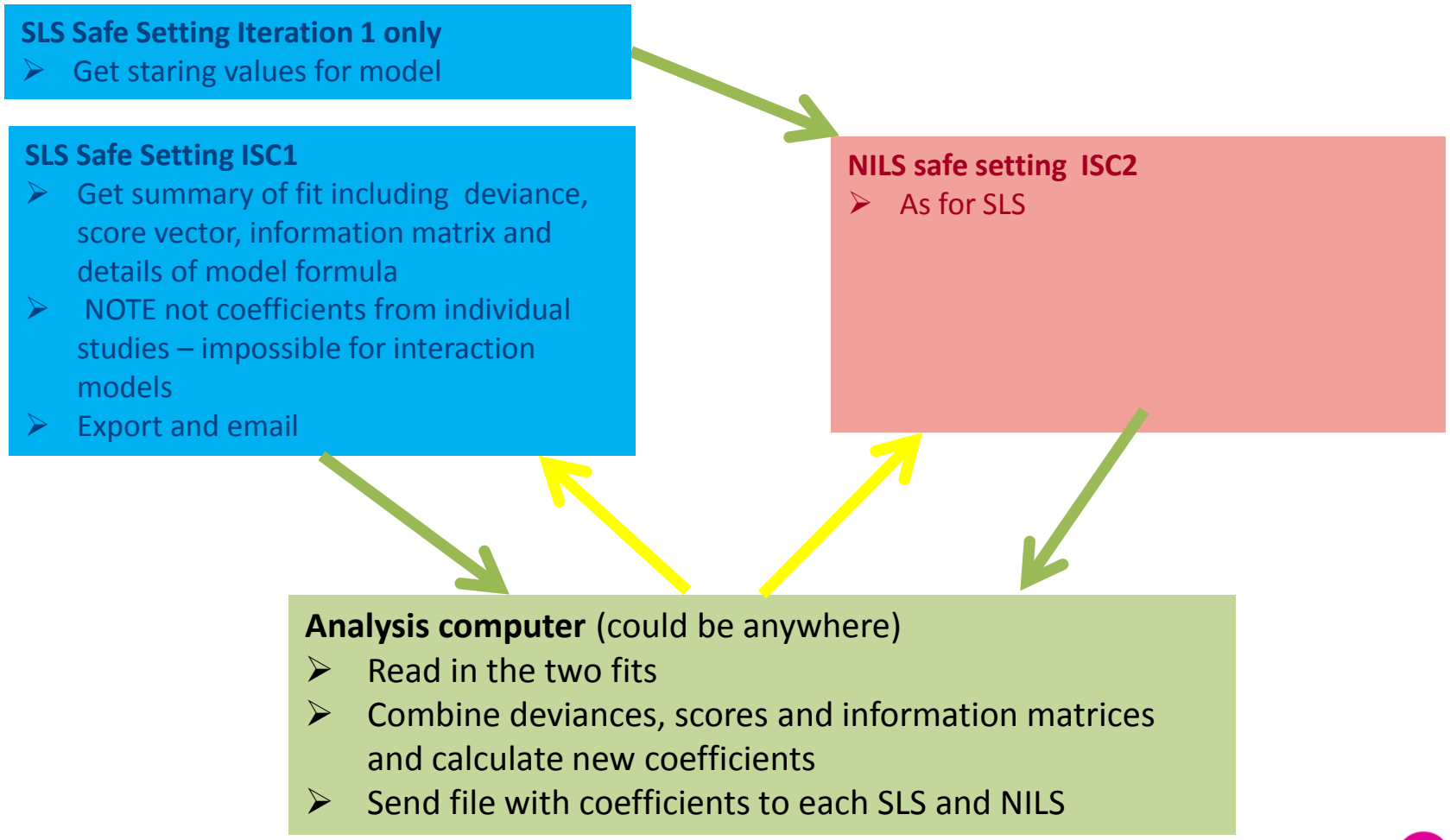
SLS safe setting
forfit_1<-eds.prepare(forfit_m,totstuds=2,studyno=1)
eds.struct(data=forfit_1,output="forfit_1_struct.R")

NILS safe setting
Same thing forfit_2

Analysis or other computer
source("forfit_1_struct.R")
source("forfit_2_struct.R")
eds.compare(struct_forfit_1,struct_forfit_2)
Hope to get output like this
Structures struct_forfit_1 and struct_forfit_2 compared
All names of 2 sources agree

Comparing levels for common variables
Levels agree Age
Levels agree econpo9
Levels agree Hed
Levels agree lex.dur (offset follow up time for the poisson model)
Levels agree lex.Xst (no of events for the poisson model)
Levels agree sclas9
Levels agree studyno

Fitting the models- iterate until convergence



Sample of exchanged file

fit_forfit_2<-structure(list(itno = 1, info = structure(c(1475146.91666667, 2142.1666666667, 78843.8333333333, 129523.083333333, 168371.333333333, 169939.166666667, 159459, 147526.25, 137003.666666667, 125500.75, 115934.166666667, 1475146.91666667, 28142.1666666667, 78843.8333333333, 129523.083333333, 168371.333333333, 169939.166666667, 159459, 147526.25, 137003.666666667, 125500.75, 115934.166666667, 28142.1666666667, 28142.1666666667, 0, 0, 0, 0, 0, 0, 0, 0, 0, 28142.1666666667, 28142.1666666667, 0, 0, 0, 0, 0, 0, 0, 0, 0, 78843.8333333333, 0, 78843.8333333333, 0, 0, 0, 0, 0, 0, 0, 0, 78843.8333333333, 0, 78843.8333333333, 0, 0, 0, 0, 0, 0, 0, 0, 129523.083333333, 0, 0, 129523.083333333, 0, 0, 0, 0, 0, 0, 0, 0, 129523.083333333, 0, 0, 129523.083333333, 0, 0, 0, 0, 0, 0, 0, 0, 168371.333333333, 0, 0, 0, 168371.333333333, 0, 0, 0, 0, 0, 0, 168371.333333333, 0, 0, 0, 168371.333333333, 0, 0, 0, 0, 0, 0, 169939.166666667, 0, 0, 0, 0, 169939.166666667, 0, 0, 0, 0, 0, 169939.166666667, 0, 0, 0, 0, 169939.166666667, 0, 0, 0, 0, 0, 0, 159459, 0, 0, 0, 0, 0, 159459, 0, 0, 0, 0, 159459, 0, 0, 0, 0, 0, 159459, 0, 0, 0, 0, 147526.25, 0, 0, 0, 0, 0, 0, 147526.25, 0, 0, 0, 147526.25, 0, 0, 0, 0, 0, 0, 147526.25, 0, 0, 0, 137003.666666667, 0, 0,
0, 0, 0, 0, 0, 137003.666666667, 0, 0, 137003.666666667, 0, 0, 0, 0, 0, 0, 0, 137003.666666667, 0, 0, 125500.75, 0, 0, 0, 0, 0, 0, 0, 0, 125500.75, 0, 125500.75, 0, 0, 0, 0, 0, 0, 0, 0, 125500.75, 0, 115934.166666667, 0, 0, 0, 0, 0, 0, 0, 0, 0, 115934.166666667, 115934.166666667, 0, 0, 0, 0, 0, 0, 0, 0, 0, 115934.166666667, 1475146.91666667, 28142.1666666667, 78843.8333333333, 129523.083333333, 168371.333333333, 169939.166666667, 159459, 147526.25, 137003.666666667, 125500.75, 115934.166666667, 1475146.91666667, 28142.1666666667, 78843.8333333333, 129523.083333333, 168371.333333333, 169939.166666667, 159459, 147526.25, 137003.666666667, 125500.75, 115934.166666667, 28142.1666666667, 28142.1666666667, 0, 0, 0, 0, 0, 0, 0, 0, 0, 28142.1666666667, 28142.1666666667, 0, 0, 0, 0, 0, 0, 0, 0, 0, 78843.8333333333, 0, 78843.8333333333, 0, 0, 0, 0, 0, 0, 0, 0, 78843.8333333333, 0, 78843.8333333333, 0, 0, 0, 0, 0, 0, 0, 0, 129523.083333333, 0, 0, 129523.083333333, 0, 0, 0, 0, 0, 0, 0, 129523.083333333, 0, 0, 129523.083333333, 0, 0, 0, 0, 0, 0, 0, 0, 168371.333333333, 0, 0, 0, 68371.333333333, 0, 0, 0, 0, 0, 0, 168371.333333333, 0, 0, 0, 168371.333333333, 0, 0, 0, 0, 0, 0, 169939.166666667, 0, 0, 0, 0, 169939.166666667, 0, 0, 0, 0, 0, 169939.166666667, 0, 0, 0, 0, 169939.166666667, 0, 0, 0, 0, 0, 0, 159459, 0, 0, 0, 0, 0, 159459, 0, 0, 0, 0, 159459, 0, 0, 0, 0, 0, 159459, 0, 0, 0, 0, 147526.25, 0, 0, 0, 0, 0, 0, 147526.25, 0, 0, 0, 147526.25, 0, 0, 0, 0, 0, 0, 147526.25, 0, 0, 0, 137003.666666667, 0, 0, 0, 0, 0, 0, 0, 137003.666666667, 0, 0, 137003.666666667, 0, 0, 0, 0, 0, 0, 0, 137003.666666667, 0, 0, 125500.75, 0, 0, 0, 0, 0, 0, 0, 0, 125500.75, 0, 125500.75, 0, 0, 0, 0, 0, 0, 0, 0, 125500.75, 0, 115934.166666667, 0, 0, 0, 0, 0, 0, 0, 0, 0, 115934.166666667, 115934.166666667, 0, 0, 0, 0, 0, 0, 0, 0, 0, 115934.166666667), .Dim = c(22L, 22L), .Dimnames = list(c("(Intercept)", "Age1", "Age2", "Age3", "Age4", "Age5", "Age6", "Age7", "Age8", "Age9", "Age10", "studyno2", "Age1:studyno2", "Age2:studyno2", "Age3:studyno2", "Age4:studyno2", "Age5:studyno2", "Age6:studyno2", "Age7:studyno2", "Age8:studyno2", "Age9:studyno2", "Age10:studyno2"), c("(Intercept)", "Age1", "Age2", "Age3", "Age4", "Age5", "Age6", "Age7", "Age8", "Age9", "Age10", "studyno2", "Age1:studyno2", "Age2:studyno2", "Age3:studyno2", "Age4:studyno2", "Age5:studyno2", "Age6:studyno2", "Age7:studyno2", Age8:studyno2", "Age9:studyno2", "Age10:studyno2"))), score = structure(c(-1459422.91666667, -28134.1666666667, -78811.8333333333, -129424.083333333, -68216.333333333, -169664.166666667, -159082, -146949.25, -136177.666666667, -124193.75, -113812.166666667, -1459422.91666667, -28134.1666666667, -8811.8333333333, -129424.083333333, -168216.333333333, -169664.166666667, -159082, -146949.25, -136177.666666667, -124193.75, -113812.166666667), .Dim = 22L, 1L), .Dimnames = list(c("(Intercept)", "Age1", "Age2", "Age3", "Age4", "Age5", "Age6", "Age7", "Age8", "Age9", "Age10", "studyno2", "Age1:studyno2", Age2:studyno2", "Age3:studyno2", "Age4:studyno2", "Age5:studyno2", "Age6:studyno2", "Age7:studyno2", "Age8:studyno2", "Age9:studyno2", "Age10:studyno2"), ULL)), start = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), dev = 2804645.67204873, devold = Inf, formula = c("~", "lex.Xst", "Age + studyno + Age * tudyno")), .Names = c("itno", "info", "score", "start", "dev", "devold", "formula"))

Difficulties encountered

- Major obstacle was the processes of getting files in and out of safe settings
  - Data had to be encrypted
  - Only Agency staff could take files in and out
  - Difficult to keep track of file names and models
  - Difficult to make sure you send correct files
- But we did get some initial results
- Revised routines
  - Fit a whole series of models together
  - Automate the process of defining file names from within the program
  - Build in checks so that correct data and files are being used at each iteration
  - Some basic documentation provided
- New project with revised data was run in December 2011
- Staff changes at NILS prevented it going forward
- But results have now been analysed and we are about to restart
  - Some very brief results follow
- Two other projects using the methods are now on their way

20 models fitted

male_v1_result$formulae

```
lex.Xst ~ Age
lex.Xst ~ Age + studyno
lex.Xst ~ Age + relig
lex.Xst ~ Age + relig + studyno
lex.Xst ~ Age + relig + studyno + relig * studyno
lex.Xst ~ Age + socnssec1_ + relig + studyno
lex.Xst ~ Age + ec0_ + relig + studyno
lex.Xst ~ Age + tn3_ + relig + studyno
lex.Xst ~ Age + cr3_ + relig + studyno
lex.Xst ~ Age + hlqp0 + relig + studyno
lex.Xst ~ Age + socnssec1_ + relig * studyno
lex.Xst ~ Age + ec0_ + relig * studyno
lex.Xst ~ Age + tn3_ + relig * studyno
lex.Xst ~ Age + cr3_ + relig * studyno
lex.Xst ~ Age + hlqp0 + relig * studyno
lex.Xst ~ Age + socnssec1_ + ec0_ + tn3_ + cr3_ + hlqp0
lex.Xst ~ Age + socnssec1_ + ec0_ + tn3_ + cr3_ + hlqp0 + studyno
lex.Xst ~ Age + socnssec1_ + ec0_ + tn3_ + cr3_ + hlqp0 + relig
lex.Xst ~ Age + socnssec1_ + ec0_ + tn3_ + cr3_ + hlqp0 + relig +  studyno
lex.Xst ~ Age + socnssec1_ + ec0_ + tn3_ + cr3_ + hlqp0 + relig +  relig * studyno
```

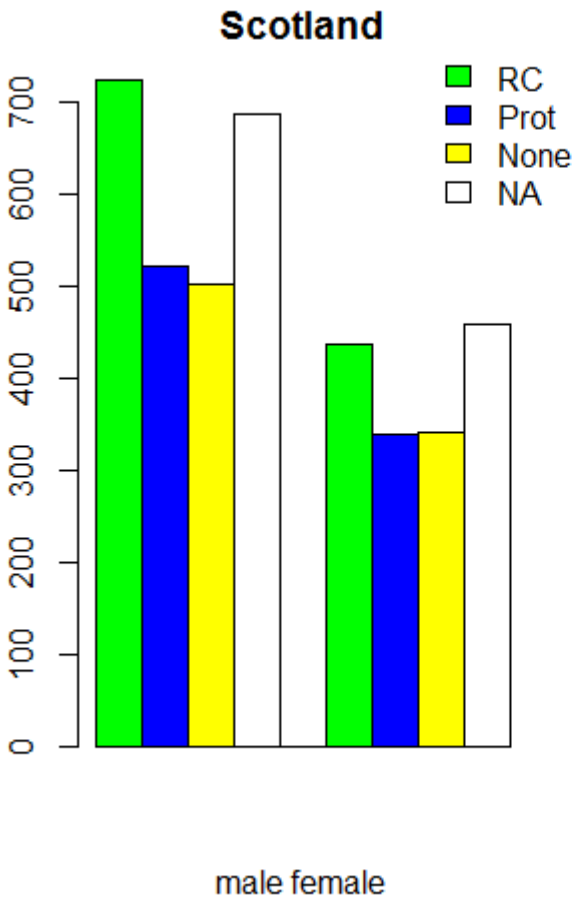
Result of joint fit can be analysed anywhere

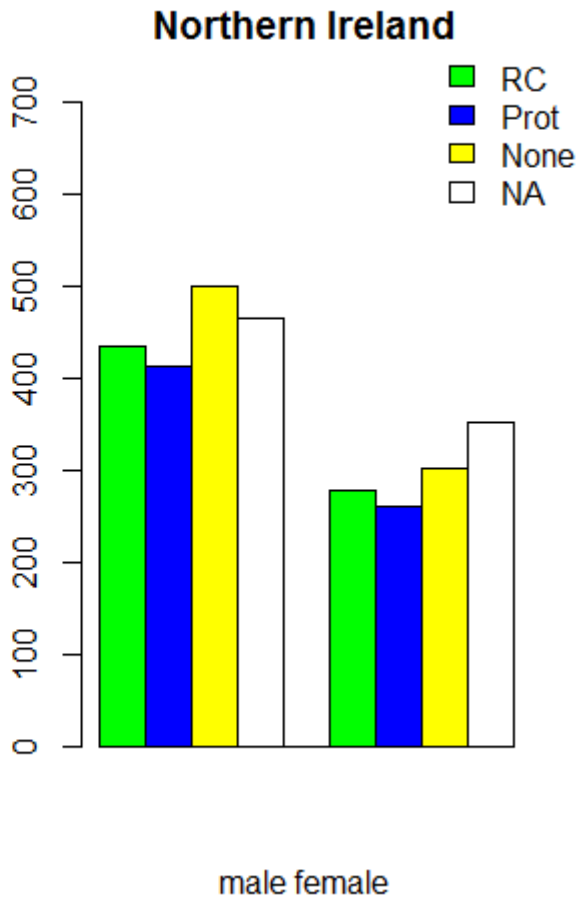
```
eds.deviance.table(male_v1_result)

 Deviance params
base                 43451.20     10
base_studyno         43323.29     11
base_relig           43431.49     13
base_studyno_relig   43288.33     14
base_studyno*relig   43239.77     17
socnssec1_           42668.43     20
eco_                 41136.48     20
tn3_                 42596.57     16
cr3_                 41737.66     16
hlqp0                42832.39     18
socnssec1_int        42621.82     23
eco_int              41092.29     23
tn3_int              42554.26     19
cr3_int              41714.23     19
hlqp0_int            42788.07     21
all                  40393.95     30
all_studyno          40279.06     31
all_relig            40374.82     33
all_relig_studyno    40271.40     34
all_relig*studyno    40245.38     37
```

Details of individual models can be examined

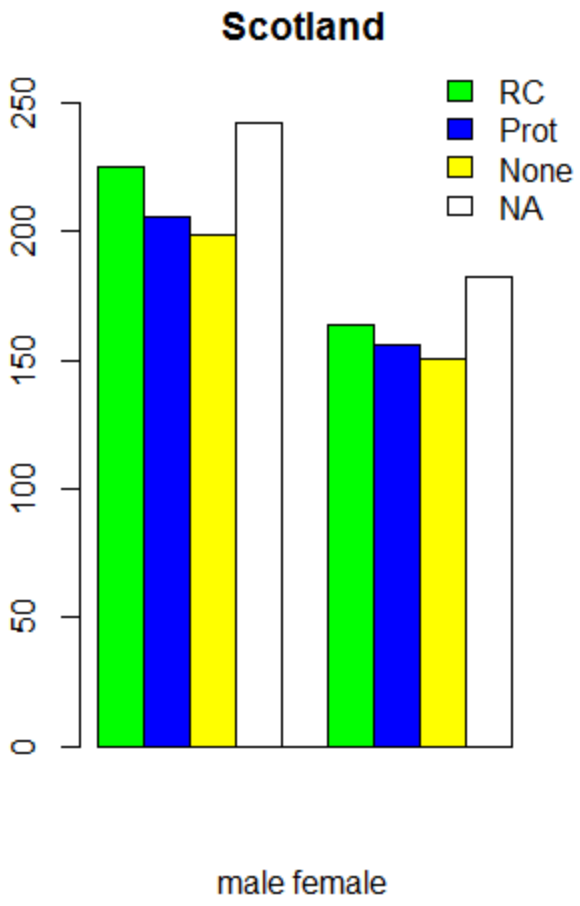
```
eds.summary(male_v1_result,5)
Results for model  base_studyno*relig
Fitted to a  poisson  model
With offset  log    lex.dur
Formula  lex.Xst ~ Age + relig + studyno + relig * studyno
                                          coef     se       z   pval
(Intercept)                            -3.3508 0.0267 -125.33 0.0000
Age25                                  -3.7522 0.1808  -20.75 0.0000
Age30                                  -3.7885 0.1052  -36.02 0.0000
Age35                                  -3.3504 0.0779  -43.02 0.0000
Age40                                  -2.9588 0.0637  -46.43 0.0000
Age45                                  -2.5468 0.0549  -46.40 0.0000
Age50                                  -2.0382 0.0460  -44.29 0.0000
Age55                                  -1.5215 0.0382  -39.80 0.0000
Age60                                  -0.9832 0.0334  -29.41 0.0000
Age65                                  -0.5234 0.0307  -17.07 0.0000
religCatholic  vs Prot                  0.3297 0.0434    7.59 0.0000
religNone  vs Prot                     -0.0393 0.0434   -0.90 0.3658
religNot answered  vs Prot              0.2768 0.0806    3.43 0.0006
studyno2                               -0.2327 0.0300   -7.75 0.0000
religCatholic  vs Prot:studyno2        -0.2785 0.0543   -5.13 0.0000
religNone  vs Prot:studyno2             0.2279 0.0682    3.34 0.0008
religNot answered  vs Prot:studyno2    -0.1575 0.1104   -1.43 0.1537
```

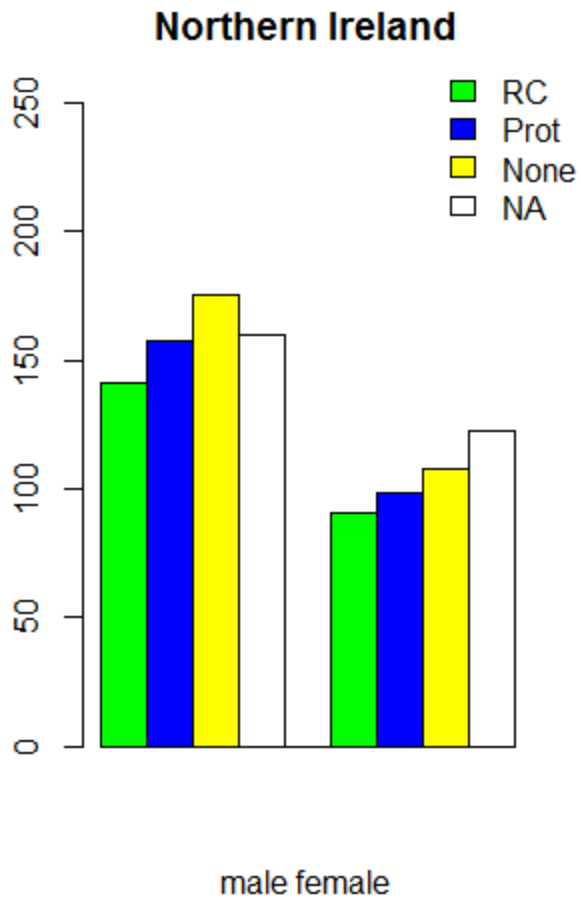
Age standardised mortality by religion





Age standardised mortality adjusted for social factors





Current status

- More work needed
- On this project to expand and check models
  - These results are very preliminary
- And on e-DataSHIELD routines
  - To make them even easier to use
  - Improve documentation
  - To create them as an R package
- We hope to get input from beyond the LSs

Acknowledgements

- Staff at NILS Andrew McCulloch, Mike Rosato, Dermot O'Reilly and others
- Staff at SLS Fiona Cox, Susan Walker and Chris Dibben and others
- DataSHIELD team for inspiration
- Our funders Scottish and NI Governments and Economic and Social Research Council (ESRC)
- Census output is Crown copyright and is reproduced with the permission of the Controller of Her Majesty's Stationery Office and the Queen's Printer for Scotland.