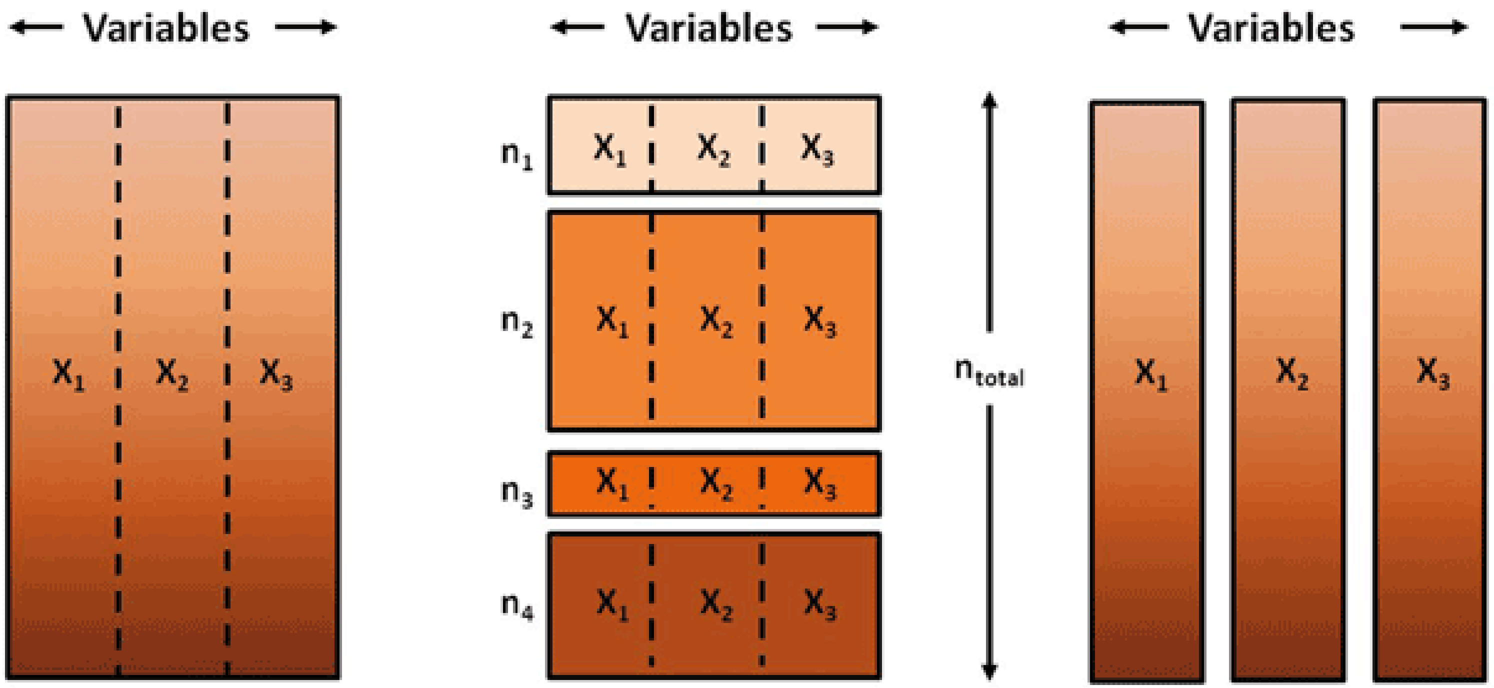
An overview of DataSHIELD

Dr Olly Butters, University of Liverpool

@DrOllyButters olly.butters@liverpool.ac.uk

Data partitioning



(a) Ideal:
All individual-level data pooled together in one file

(b) Horizontally partitioned: individual-level data held in 4 separate data files, one for each study

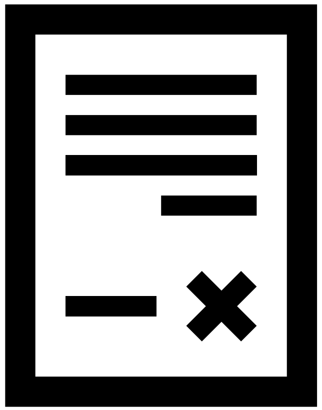
(c) Vertically partitioned : individual-level data held in 3 datasets distinct data files, one variable for each study

Gaye et al, 2014: 10.1093/ije/dyu188

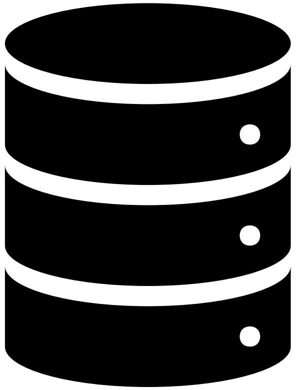
Problems collaborating

Data sharing-access barriers result from a range of scenarios

Ethico-legal & governance



Control IP



Physical size of data

@DrOllyButters olly.butters@liverpool.ac.uk

Disclosure Control

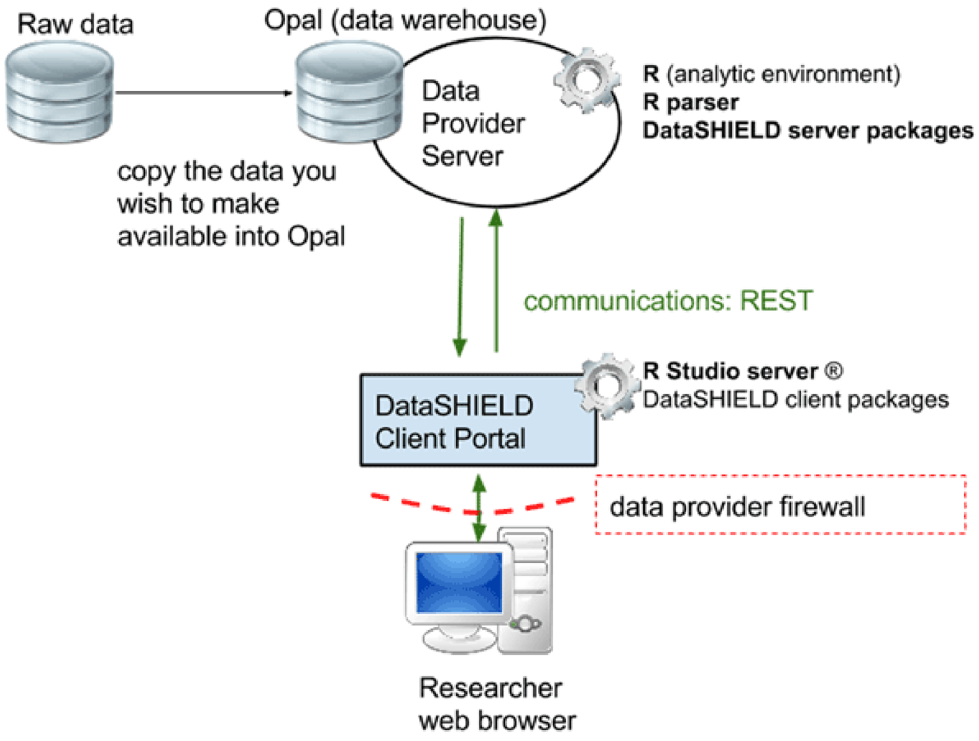
- Methods to prevent disclosure of individual level data
- Typically in medical and health research:
  - Anonymisation
  - Psuedonymisation
  - Use of a data safe haven
    - Log onto a portal can view all the individual level data
  - Reliant on data governance measures
    - Sign a contract saying you won't misuse the data
    - Are any penalties enforced?
  - For very sensitive information human scrutiny of outputs
    - Financial and time cost

UNIVERSITY OF LIVERPOOL

DataSHIELD: a solution

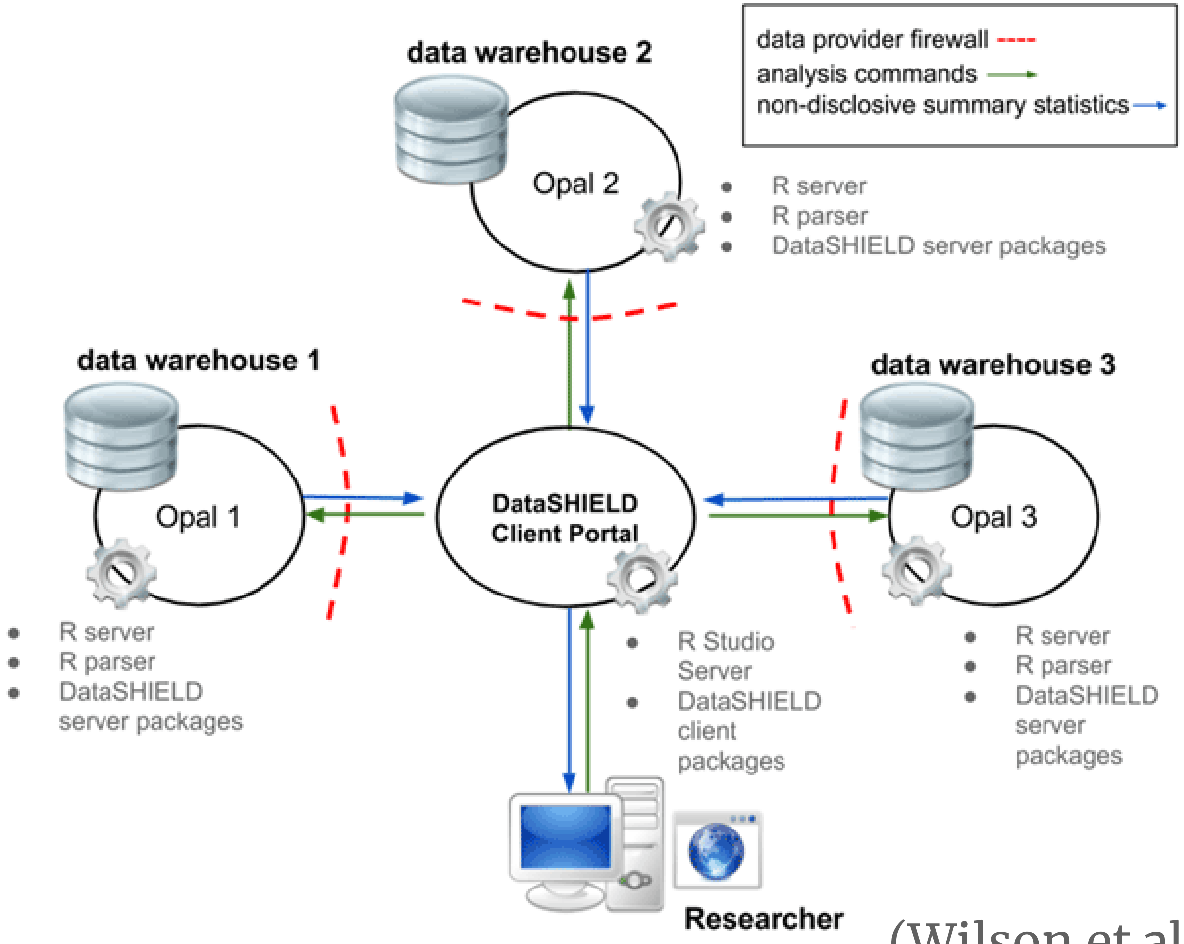
- Technological and methodological solution
- Designed by epidemiologists and biostatisticians, community led
- Individual level and raw data do not leave the data controller site
    - Simplifies and streamlines data governance process and decisions
- Automated disclosure control built into the system
    - Addresses confidentiality restrictions
    - Only non-disclosive summary statistics returned to analyst
    - No delay waiting for human scrutiny of results
    - GDPR compliance
- Analyse data from multiple studies (meta-analysis) or a single one
- Researcher does their own analysis in real time
- Open source – free at point of use (cost effective)

Single-site DataSHIELD



(Wilson et al., 2017: [10.5334/dsj-2017-021](https://doi.org/10.5334/dsj-2017-021))

Multi-site DataSHIELD



(Wilson et al., 2017: 10.5334/dsj-2017-021)

UNIVERSITY OF LIVERPOOL

DataSHIELD functions

Coercing functions
- ds.asCharacter
- ds.asInteger
- ds.asMatrix
- ds.asDataMatrix
- ds.asList
- ds.asNumeric
- ds.asFactor
- ds.asLogical

Number manipulation functions
- ds.make
- ds.Boole
- ds.list
- ds.abs
- ds.log
- ds.exp
- ds.sqrt
- ds.rep
- ds.seq
- ds.c

- **Administrative functions**
- ds.listClientsideFunctions
- ds.listServersideFunctions
- ds.message
- ds.setSeed
- ds.listDisclosureSettings
- ds.ls
- ds.rm
- ds.testObjExists
- ds.look
- ds.listOpals
- ds.setDefaultOpals

Data Frame/List manipulation functions
- ds.completeCases
- ds.replaceNA
- ds.sample
- ds.merge
- ds.recodeValues
- ds.getWGSR
- ds.dataFrame
- ds.dataFrameSubset
- ds.dataFrameFill
- ds.dataFrameSort
- ds.tapply
- ds.tapply.assign
- ds.vectorCalc
- ds.list
- ds.unList
- ds.assign
- ds.cbind
- ds.rbind

Factor manipulation functions
- ds.changeRefGroup
- ds.recodeLevels

Matrices Functions
- ds.matrix
- ds.matrixDiag
- ds.matrixMult
- ds.matrixDet
- ds.matrixDimnames
- ds.matrixTranspose
- ds.matrixDet.report
- ds.matrixInvert

Data structure queries
- ds.levels
- ds.length
- ds.exists
- ds.colnames
- ds.names
- ds.dim
- ds.isValid
- ds.isNA
- ds.numNA
- ds.class

Summary Statistics Functions
- ds.mean
- ds.meanByClass
- ds.meanSdGp
- ds.quantileMean
- ds.rowColCalc
- ds.var
- ds.cov
- ds.cor
- ds.corTest
- ds.summary
- ds.skewness
- ds.kurtosis

Table functions
- ds.table

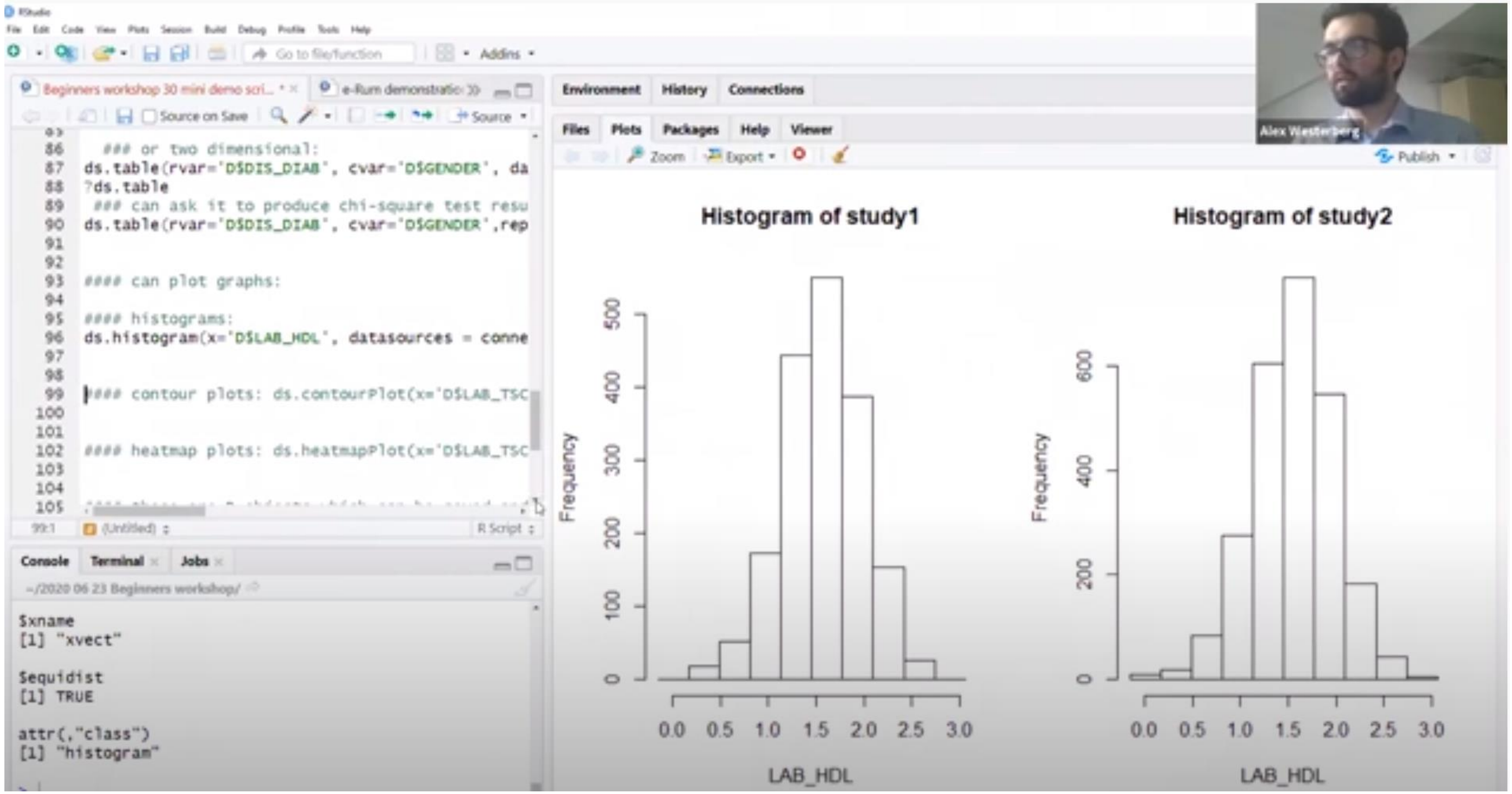
Survival Analysis functions
- ds.lexis
- ds.reShape

Distribution Generating functions
- ds.rNorm
- ds.rBinom
- ds.rPois
- ds.rUnif

Modelling Functions
- ds.glm
- ds.glmSummary
- ds.glmSLMA
- ds.glmPredict
- ds.lmerSLMA
- ds.glmerSLMA

Plotting functions
- ds.histogram
- ds.boxPlot
- ds.densityGrid
- ds.heatmapPlot
- ds.contourPlot
- ds.scatterPlot

What’s it like to use?



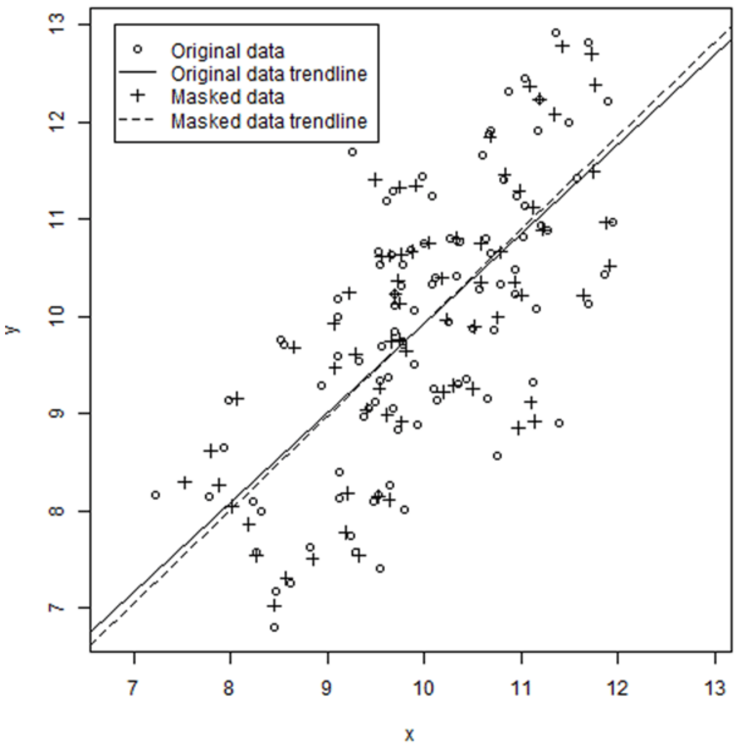
https://youtu.be/f0NJNu--Oik

UNIVERSITY OF LIVERPOOL

Statistical Disclosure Control

- Systematically built into each function – technique depends on function
- No “print to screen” of microdata
- Cell suppression
  - Minimum cell count – set by each study (typically 5 or 3)
  - Prevent disclosure of individual level data in tables, histograms etc
- Subset suppression
- GLM number of parameters limit
- String length limit
- K-nearest neighbour K value
- Number of categories in categorical data
- Amount of noise to add
- Often just adapt a standard R function
  - Disclosive information blocked from function
  - ds.glm vs glm - residuals are blocked

Data visualization in DataSHIELD



(Avraam et al., 2021: [10.1140/epjds/s13688-020-00257-4](https://doi.org/10.1140/epjds/s13688-020-00257-4))

- Scatter plots inherently disclosive
- Application of different methods in DataSHIELD
  - *k*-anonymization: preserves privacy by reducing the granularity of the data (suppression, generalization)
  - a deterministic approach: replaces individual-level observations with centroids of *k* nearest neighbours
  - a probabilistic procedure: perturbs individual attributes with addition of random stochastic noise
- Retain original data structure and features, no disclosure
- Now want to apply this to geospatial data – make use of full postcode/geolocation information, not aggregated areas

Other developments

- Other source data providers (Other DBs, flat files, connections etc)
- Genomics analysis (integration with BioConductor)
- Machine learning techniques
- Geospatial in the pipeline
- Exploring a spin out company to offer support for DataSHIELD
- About to write a grant application around disclosure control (interested? Let me know!)
- About to write a grant application about using DataSHIELD with TREs (interested? Let me know!)

UNIVERSITY OF LIVERPOOL

- www.datashield.ac.uk
- DataSHIELD Conference (Online) 10-11 November 2021