Cleaning and analysis of the SCTS database

Graeme L Hickey[1,2]; Stuart W Grant[2]; Kate McAllister[1]; Norman Stein[1]; Iain Buchan[1]; Ben Bridgewater[2]

[1]Northwest Institute of BioHealth Informatics
[2]University Hospital South Manchester

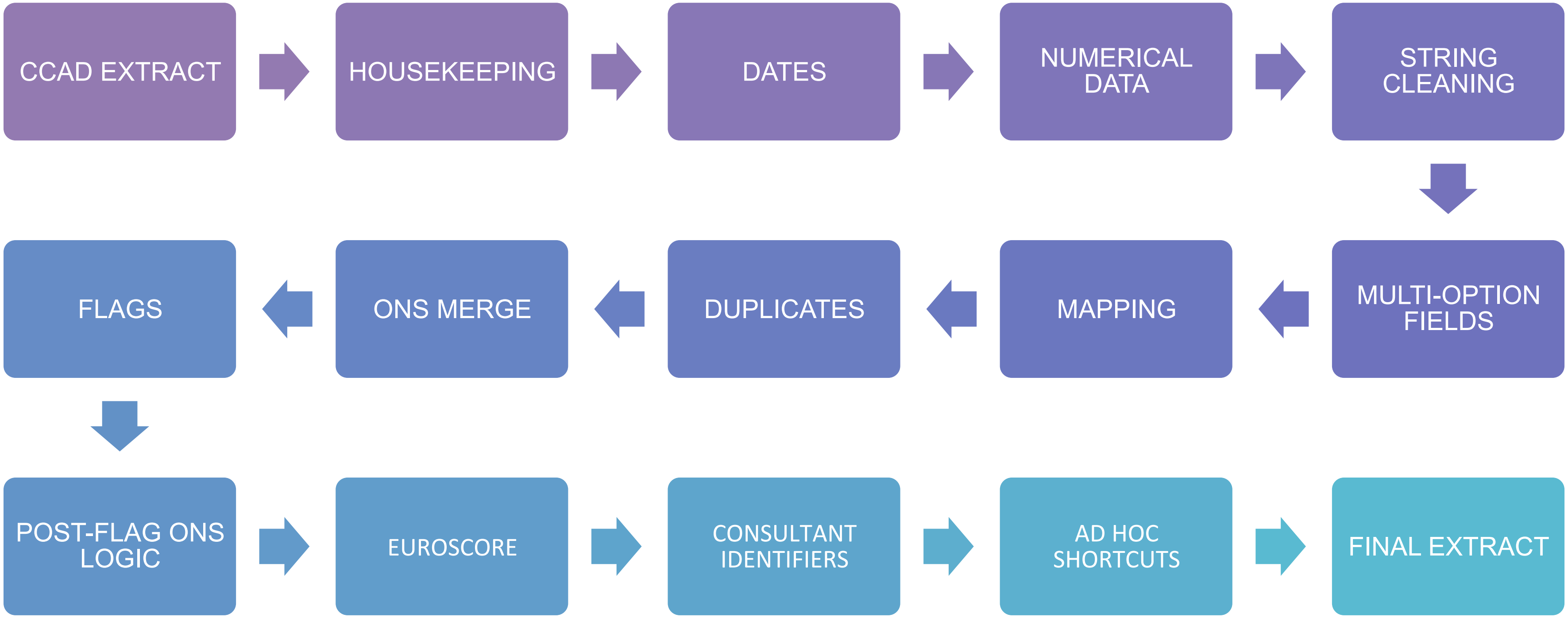
Structure: 20 March 2012

- 444,289 records pre-cleaning
- 422,493 records post-cleaning
- 181 fields made available
- 45 hospitals in UK and Ireland

- Real world data is messy:
  - missingness
  - measurement error
  - conflicts / miscoding

requires cleaning

Cleaning schema



CCAD EXTRACT

HOUSEKEEPING

DATES

NUMERICAL DATA

STRING CLEANING

MULTI-OPTION FIELDS

MAPPING

DUPLICATES

ONS MERGE

FLAGS

POST-FLAG ONS LOGIC

EUROSCORE

CONSULTANT IDENTIFIERS

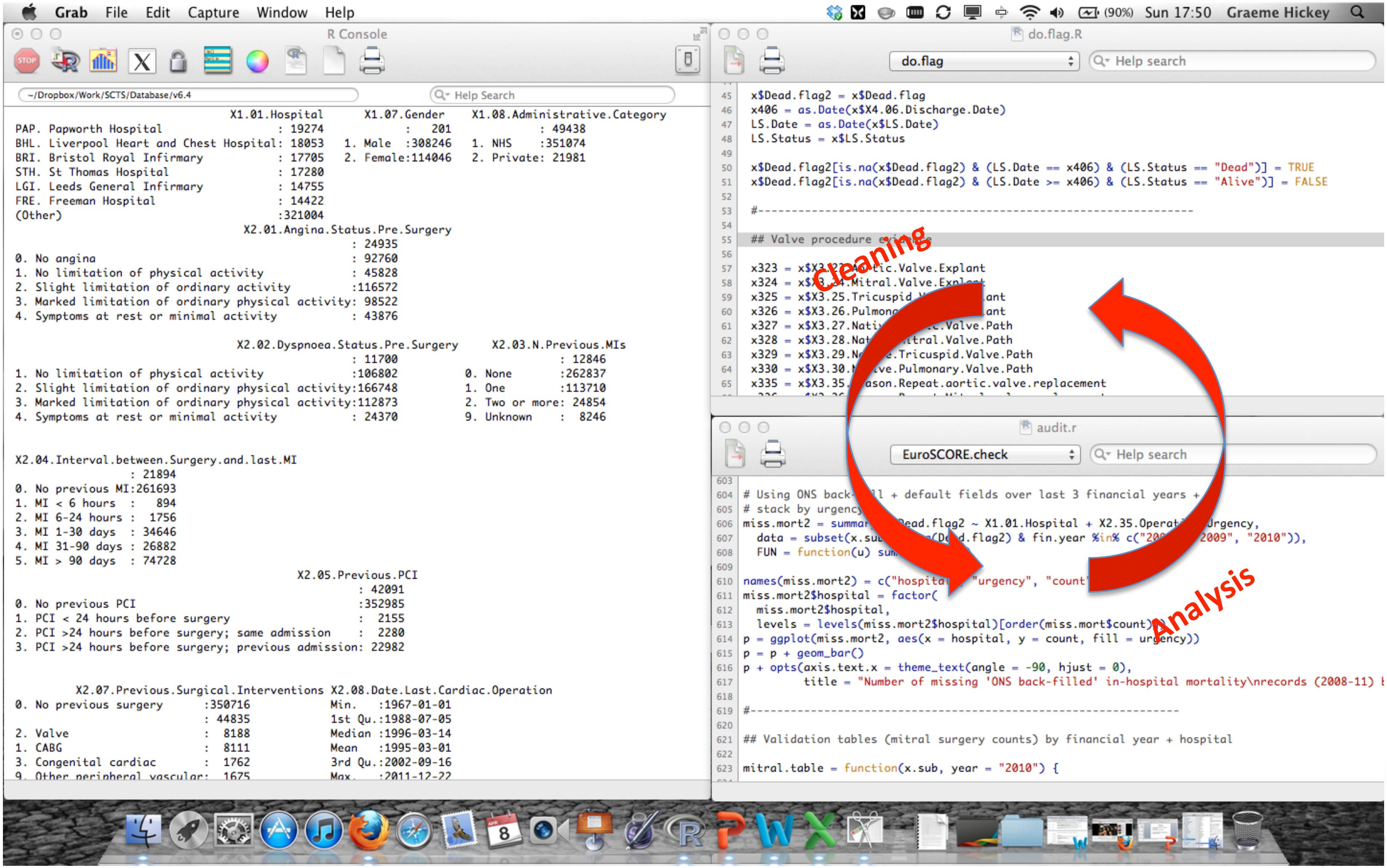
AD HOC SHORTCUTS

FINAL EXTRACT

Implementation

- R: a language and environment for statistical computing and graphics
- Transparent (common S language and open source)
- Sharable (free software);
- Reproducible (tweak and re-run)
- Programmable reports (data organisation, cleaning, analysis, presentation)
- Seamless transition from cleaning to analysis

Database in action

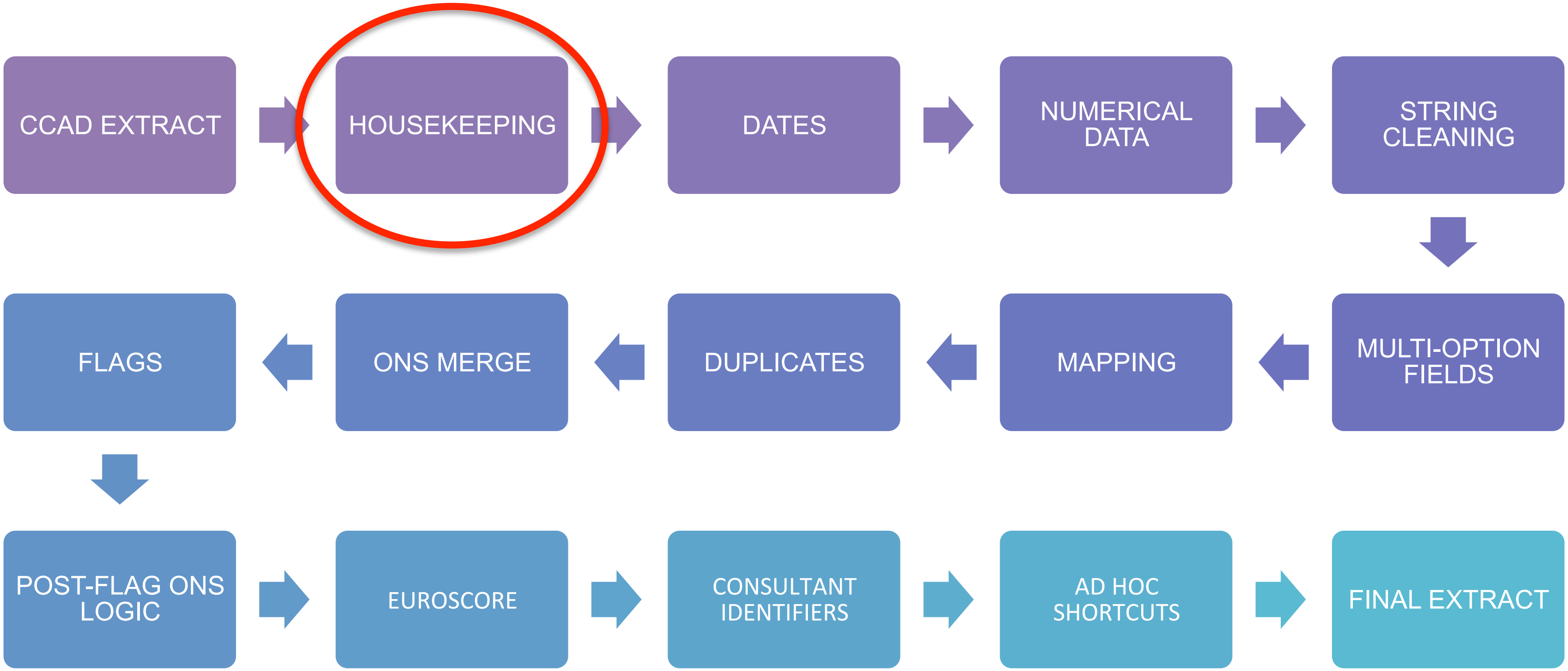

Grab File Edit Capture Window Help
(90%) Sun 17:50 Graeme Hickey
R Console
~/Dropbox/Work/SCTS/Database/v6.4
Help Search
do.flag.R
do.flag
Help search
audit.r
EuroSCORE.check
Help search
Cleaning
Analysis

```
                              X1.01.Hospital      X1.07.Gender    X1.08.Administrative.Category
PAP. Papworth Hospital                 : 19274          :   201                 : 49438
BHL. Liverpool Heart and Chest Hospital: 18053   1. Male  :308246   1. NHS     :351074
BRI. Bristol Royal Infirmary           : 17705   2. Female:114046   2. Private: 21981
STH. St Thomas Hospital                : 17280
LGI. Leeds General Infirmary           : 14755
FRE. Freeman Hospital                  : 14422
(Other)                                :321004
                                X2.01.Angina.Status.Pre.Surgery
                                                  : 24935
0. No angina                                      : 92760
1. No limitation of physical activity             : 45828
2. Slight limitation of ordinary activity         :116572
3. Marked limitation of ordinary physical activity: 98522
4. Symptoms at rest or minimal activity           : 43876

                               X2.02.Dyspnoea.Status.Pre.Surgery     X2.03.N.Previous.MIs
                                                  : 11700                        : 12846
1. No limitation of physical activity             :106802        0. None        :262837
2. Slight limitation of ordinary physical activity:166748        1. One         :113710
3. Marked limitation of ordinary physical activity:112873        2. Two or more: 24854
4. Symptoms at rest or minimal activity           : 24370        9. Unknown     :  8246


X2.04.Interval.between.Surgery.and.last.MI
                 : 21894
0. No previous MI:261693
1. MI < 6 hours  :   894
2. MI 6-24 hours :  1756
3. MI 1-30 days  : 34646
4. MI 31-90 days : 26882
5. MI > 90 days  : 74728
                                           X2.05.Previous.PCI
                                                    : 42091
0. No previous PCI                                  :352985
1. PCI < 24 hours before surgery                    :  2155
2. PCI >24 hours before surgery; same admission     :  2280
3. PCI >24 hours before surgery; previous admission: 22982


          X2.07.Previous.Surgical.Interventions X2.08.Date.Last.Cardiac.Operation
0. No previous surgery       :350716            Min.    :1967-01-01
                             : 44835            1st Qu.:1988-07-05
2. Valve                     :  8188            Median :1996-03-14
1. CABG                      :  8111            Mean   :1995-03-01
3. Congenital cardiac        :  1762            3rd Qu.:2002-09-16
9. Other peripheral vascular:  1675             Max.    :2011-12-22
```

```r
x$Dead.flag2 = x$Dead.flag
x406 = as.Date(x$X4.06.Discharge.Date)
LS.Date = as.Date(x$LS.Date)
LS.Status = x$LS.Status

x$Dead.flag2[is.na(x$Dead.flag2) & (LS.Date == x406) & (LS.Status == "Dead")] = TRUE
x$Dead.flag2[is.na(x$Dead.flag2) & (LS.Date >= x406) & (LS.Status == "Alive")] = FALSE

#-----------------------------------------------------------------

## Valve procedure evi

x323 = x$X3.2   ortic.Valve.Explant
x324 = x$  .Mitral.Valve.Expl
x325 = x$X3.25.Tricuspid        lant
x326 = x$X3.26.Pulmona          lant
x327 = x$X3.27.Nati       c.Valve.Path
x328 = x$X3.28.Nat      itral.Valve.Path
x329 = x$X3.29.N     e.Tricuspid.Valve.Path
x330 = x$X3.30.     ve.Pulmonary.Valve.Path
x335 = x$X3.35.    ason.Repeat.aortic.valve.replacement
```

```r

# Using ONS back-  ll + default fields over last 3 financial years +
# stack by urgency
miss.mort2 = summar     ead.flag2 ~ X1.01.Hospital + X2.35.Operat     Urgency,
  data = subset(x.su      (De  d.flag2) & fin.year %in% c("20      2009", "2010")),
  FUN = function(u) su

names(miss.mort2) = c("hospita    "urgency", "count"
miss.mort2$hospital = factor(
  miss.mort2$hospital,
  levels = levels(miss.mort2$hospital)[order(miss.mort$count)
p = ggplot(miss.mort2, aes(x = hospital, y = count, fill = urgency))
p = p + geom_bar()
p + opts(axis.text.x = theme_text(angle = -90, hjust = 0),
        title = "Number of missing 'ONS back-filled' in-hospital mortality\nrecords (2008-11) b

#-----------------------------------------------------------------

## Validation tables (mitral surgery counts) by financial year + hospital

mitral.table = function(x.sub, year = "2010") {
```

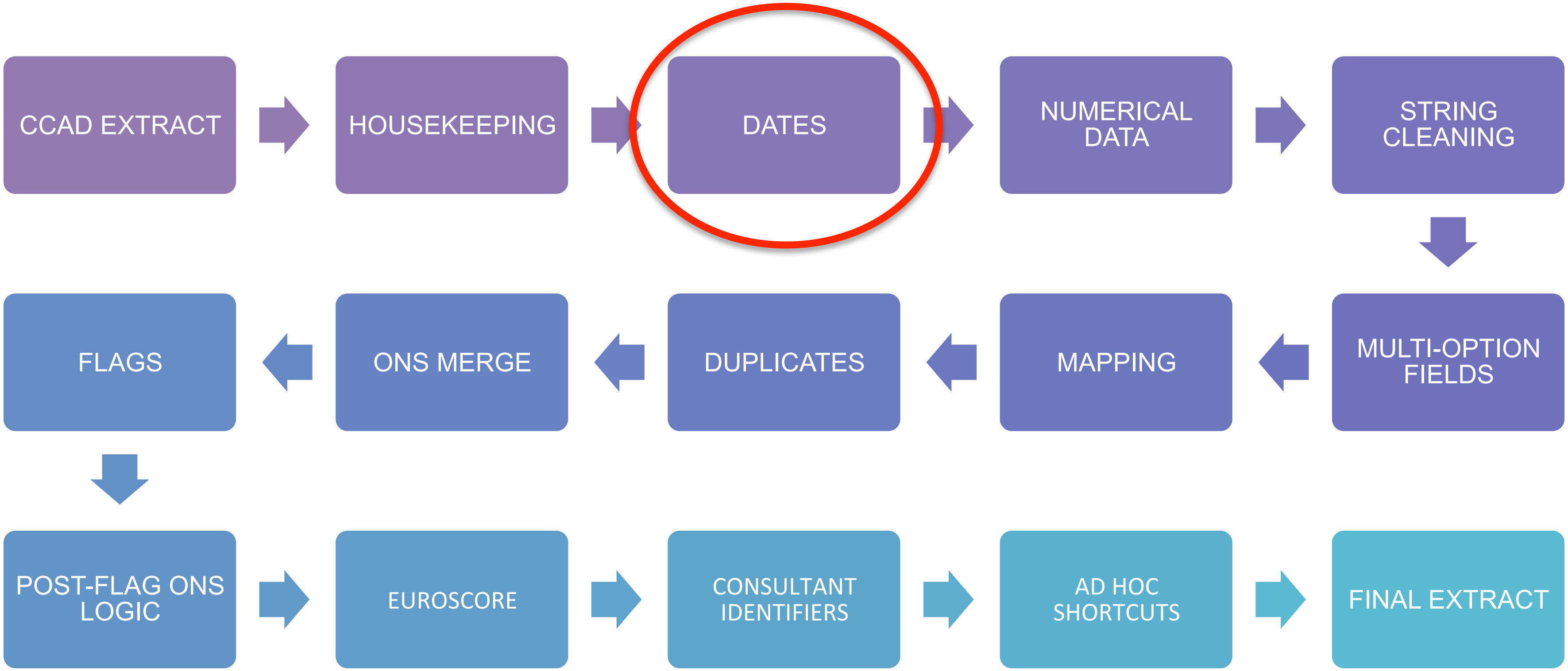
Cleaning schema



Housekeeping

- Remove identifiable fields
- Delete free text and low-importance fields
- Tidy-up field names (spelling, whitespace, etc.)

Cleaning schema

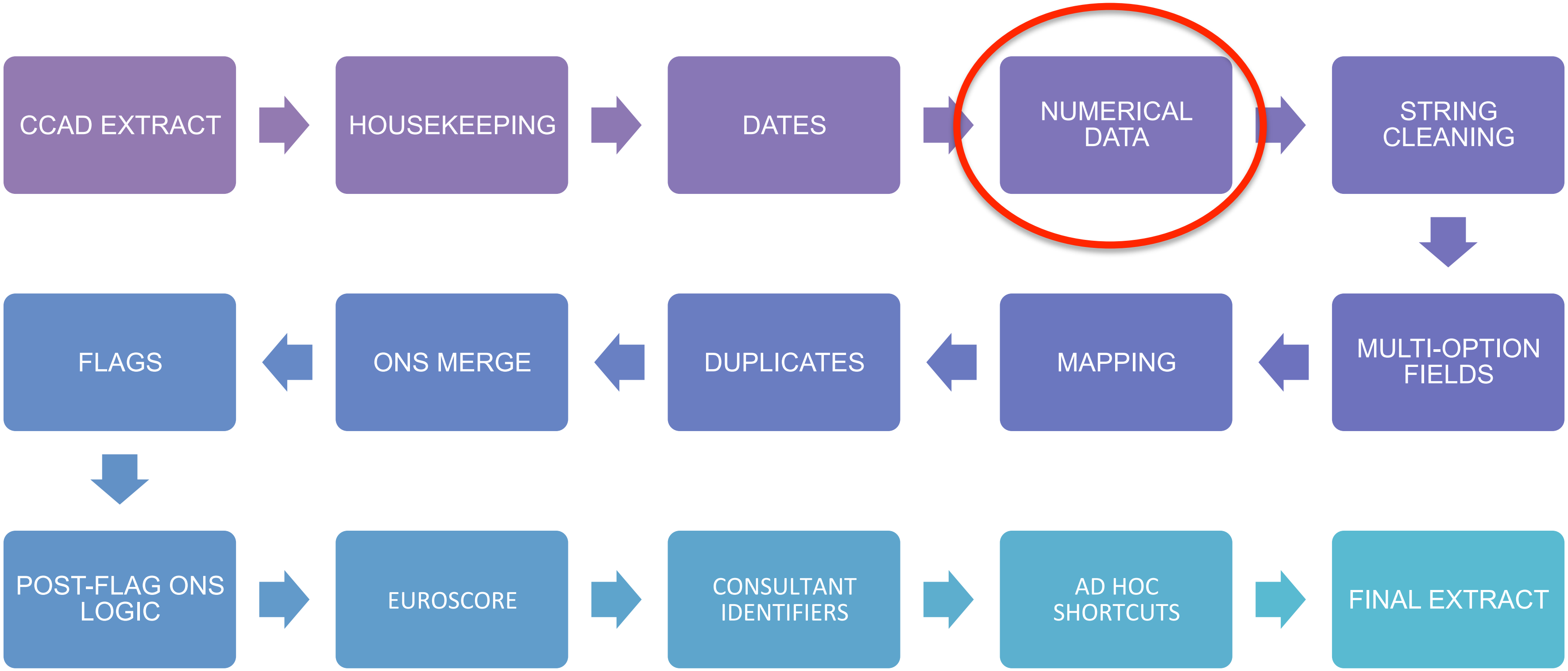


Dates

- Formatting – time discarded except for procedure
- Delete records < 1st Jan 1998
- Delete dates (pre-67 and future)
- Delete records not satisfying sensible logic:

admission ≤ procedure ≤ discharge

Cleaning schema



CCAD EXTRACT

HOUSEKEEPING

DATES

NUMERICAL DATA

STRING CLEANING

MULTI-OPTION FIELDS

MAPPING

DUPLICATES

ONS MERGE

FLAGS

POST-FLAG ONS LOGIC

EUROSCORE

CONSULTANT IDENTIFIERS

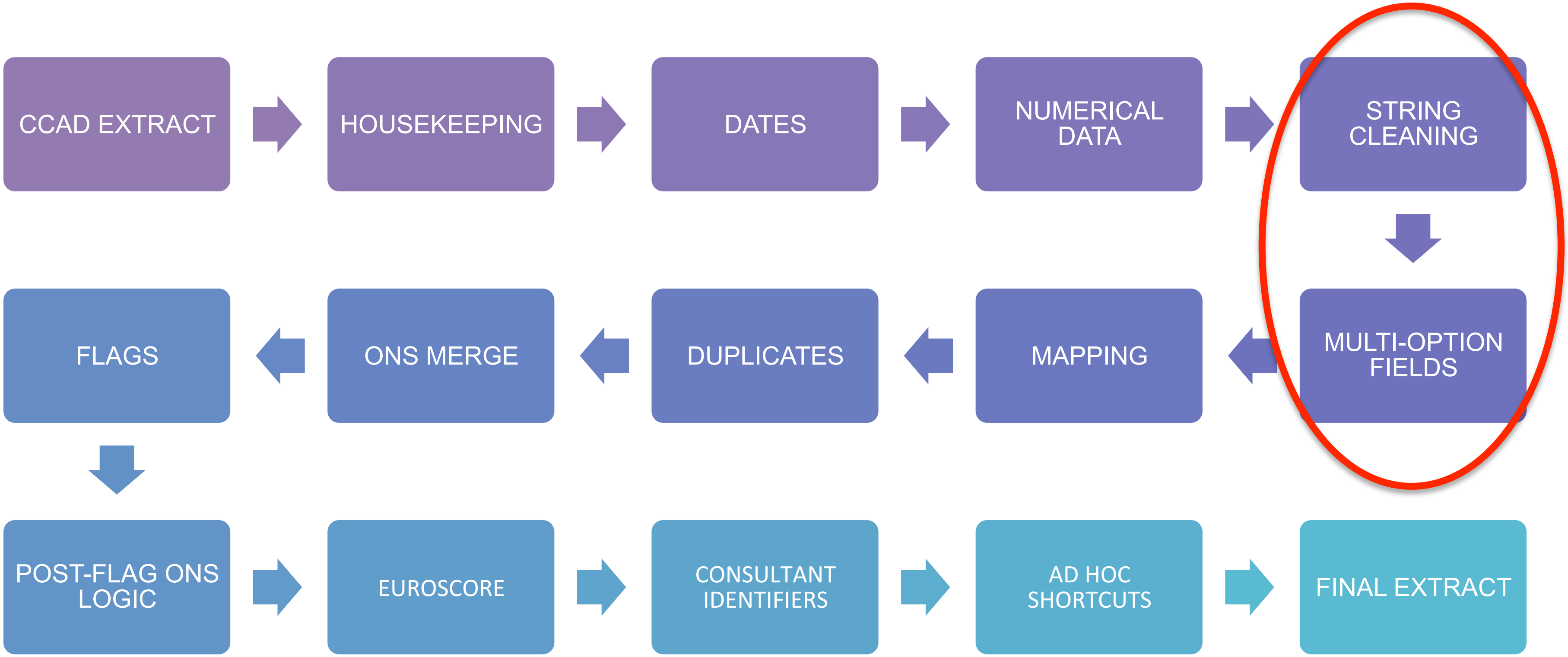
AD HOC SHORTCUTS

FINAL EXTRACT

Numerical data

- Delete free text and symbols
- Delete impossible values (e.g. 5 valves operated on)
- Delete [clinically] unlikely values (e.g. > 11 grafts)
- Resolve ‘obvious’ serial imputation errors (e.g. height recorded in mm and not cm)

Cleaning schema



CCAD EXTRACT

HOUSEKEEPING

DATES

NUMERICAL DATA

STRING CLEANING

MULTI-OPTION FIELDS

MAPPING

DUPLICATES

ONS MERGE

FLAGS

POST-FLAG ONS LOGIC

EUROSCORE

CONSULTANT IDENTIFIERS

AD HOC SHORTCUTS

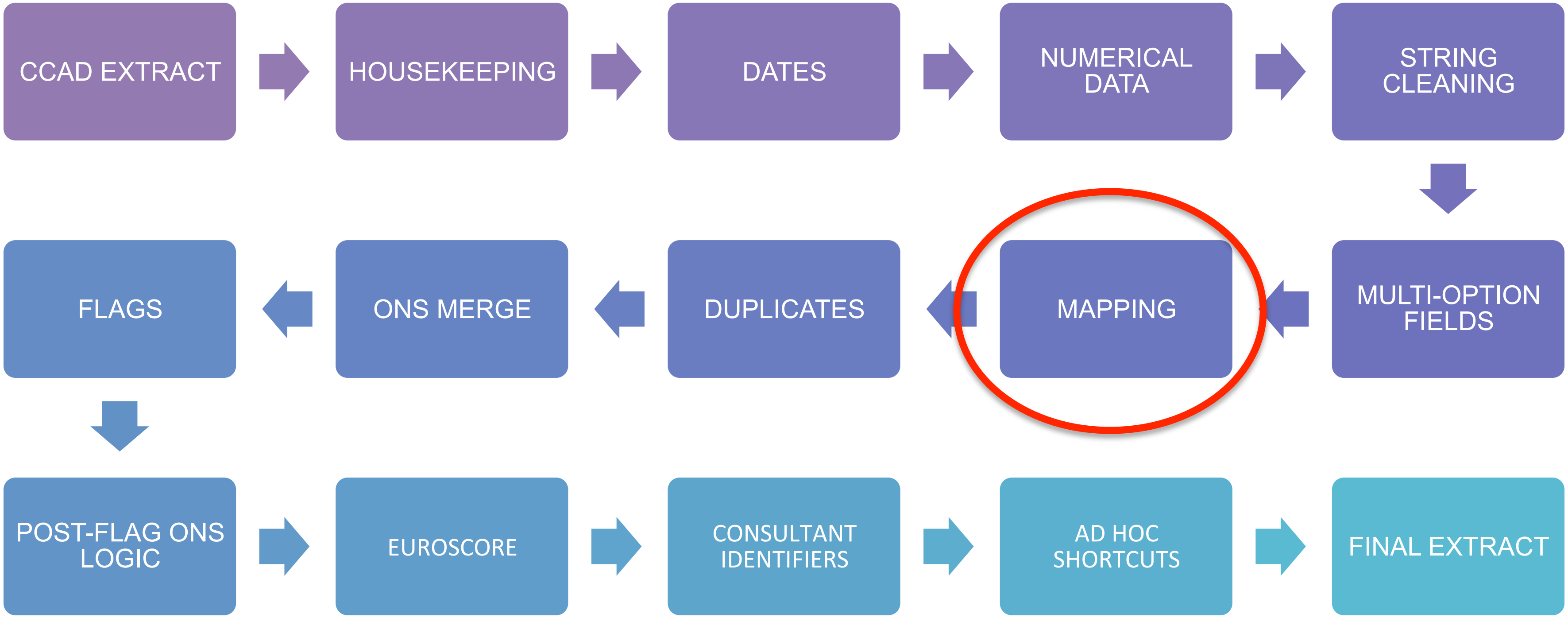
FINAL EXTRACT

NIBHI

String cleaning

- Transcriptional errors harmonized (e.g. ‘female’ ➔ ‘2. Female’)
  - manual
  - automated macros
- Invalid inputs (e.g. free text) assigned to [clinically] appropriate options
- Multi-option fields (ordered + unordered) – structure retained
- Small number of conflicts and mappings handled

Cleaning schema



CCAD EXTRACT

HOUSEKEEPING

DATES

NUMERICAL DATA

STRING CLEANING

MULTI-OPTION FIELDS

MAPPING

DUPLICATES

ONS MERGE

FLAGS

POST-FLAG ONS LOGIC

EUROSCORE

CONSULTANT IDENTIFIERS

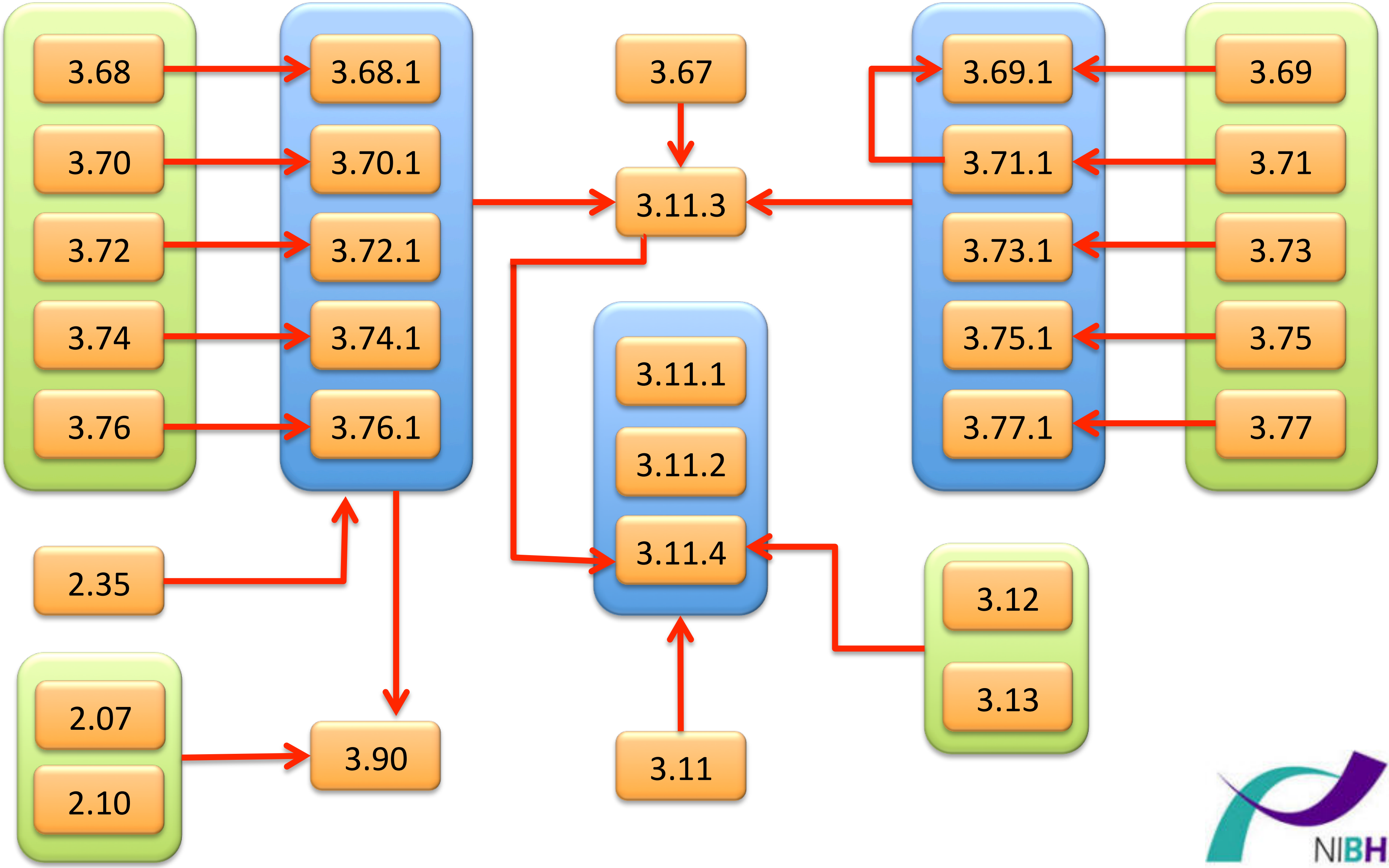
AD HOC SHORTCUTS

FINAL EXTRACT

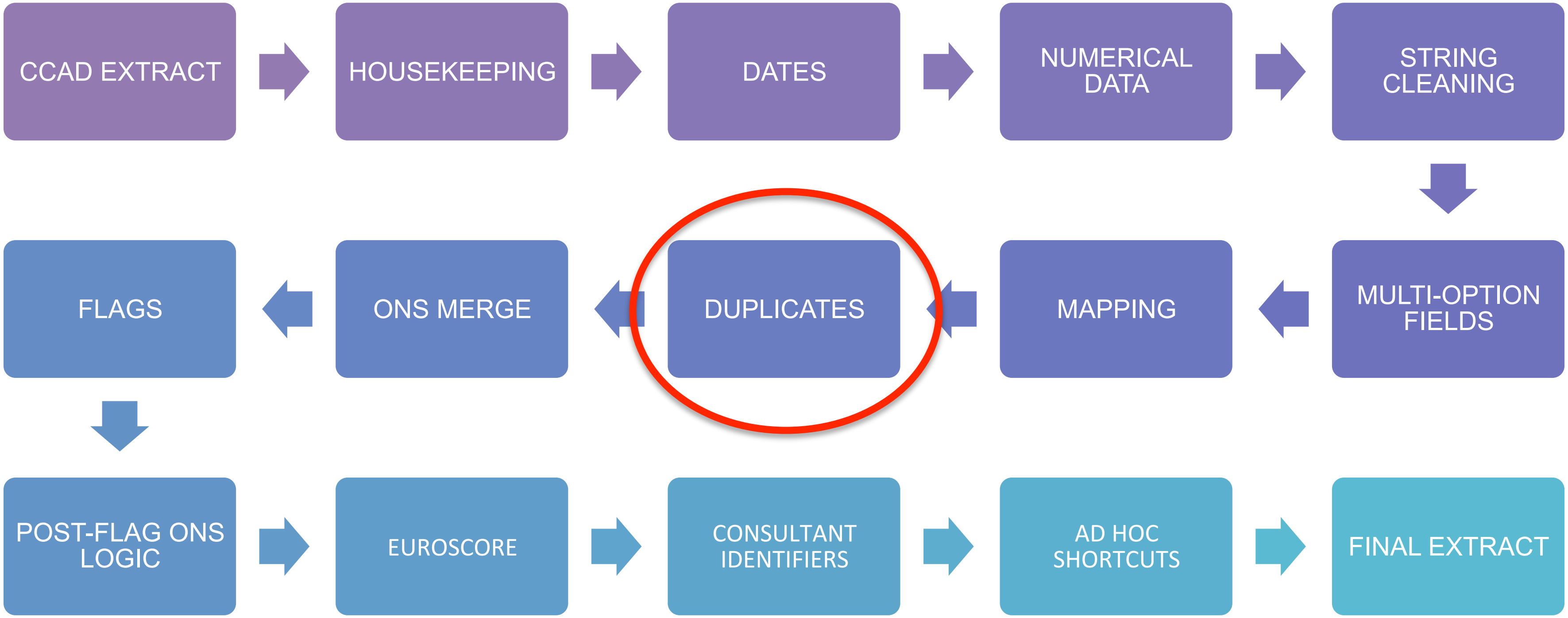
Mapping

- Partially fragmented about March 2010: Version 3 & 4.
- Scripts written to map V3.8 into V4.1.2
- Simultaneous pre- and post-mapping cleaning
- Retrospectively deleted isolated abdominal procedure records

Example: major aortic fields



Cleaning schema



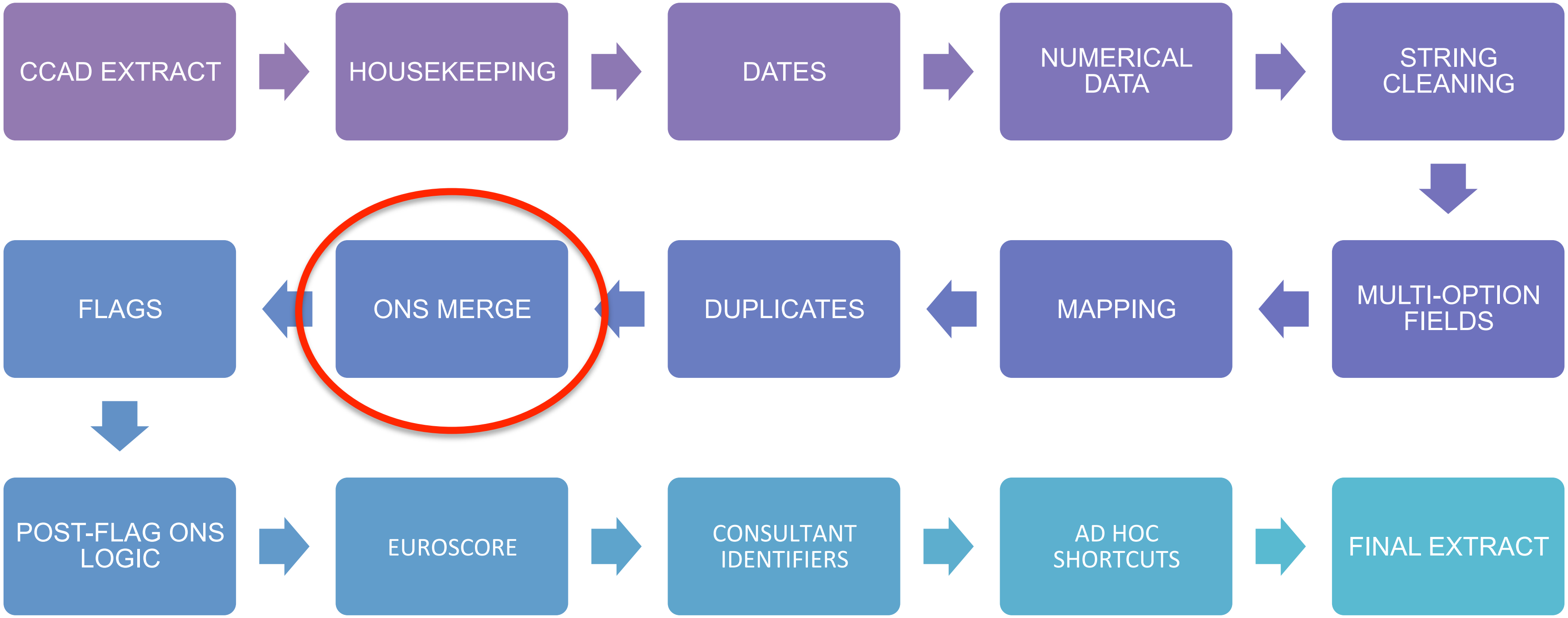
Duplicate records

- A record is classed as a duplicate if it matches on a subset.
- The most recent record created is kept; others deleted
- Records inspected after removal to 'confirm' duplicates and not re-dos

Match criteria

- ✓ hospital
- ✓ gender
- ✓ age (decimal precision)
- ✓ Apollo number (where available)
- ✓ number of previous heart operations
- ✓ procedure indicators (CABG, valve, major aortic, other)
- ✓ admission, procedure (incl. time) and discharge date
- ✓ elective (true/false)

Cleaning schema

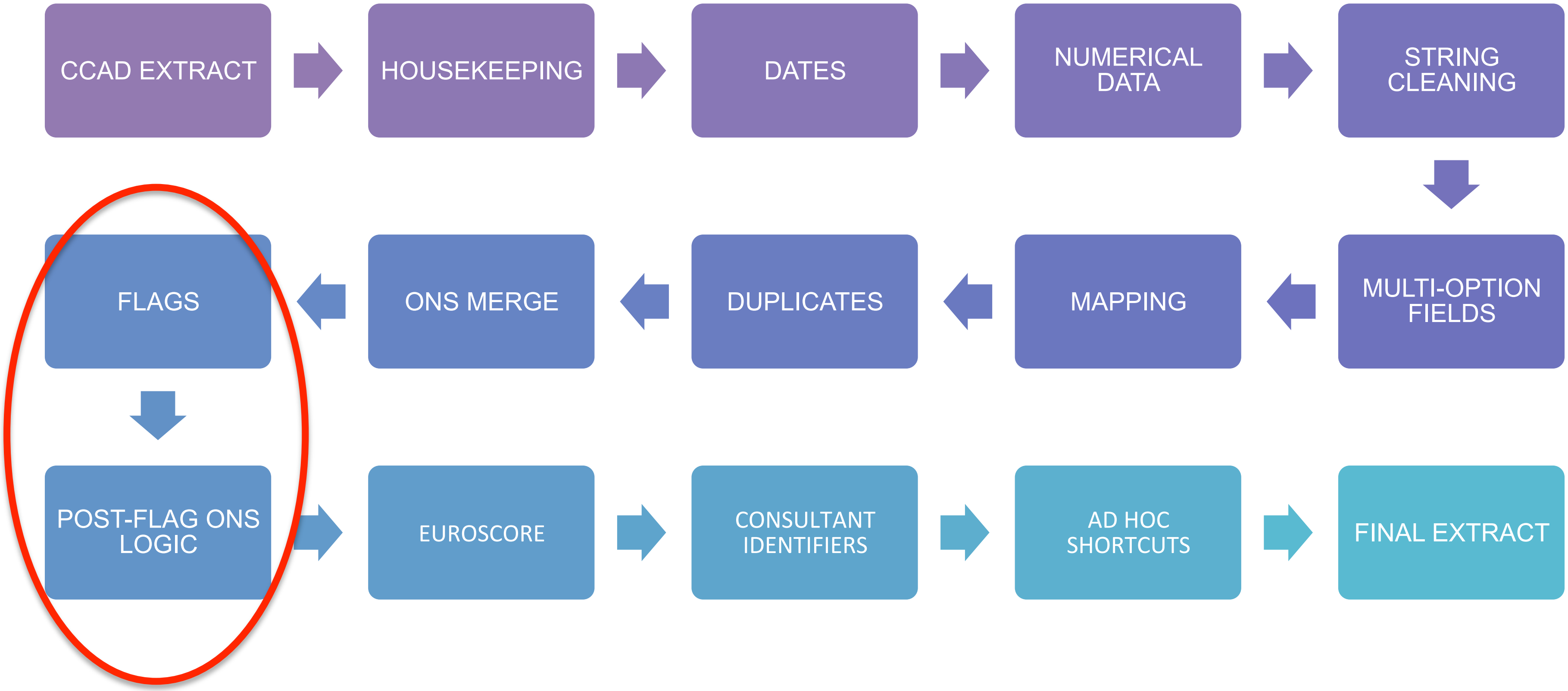


NIBHI

ONS data linkage

- Life status data extracted from the Office for National Statistics (ONS)
- ONS data removed if precedes procedure date
- Records deleted if patient deceased prior to a first-time cardiac procedure

Cleaning schema



CCAD EXTRACT

HOUSEKEEPING

DATES

NUMERICAL DATA

STRING CLEANING

MULTI-OPTION FIELDS

MAPPING

DUPLICATES

ONS MERGE

FLAGS

POST-FLAG ONS LOGIC

EUROSCORE

CONSULTANT IDENTIFIERS

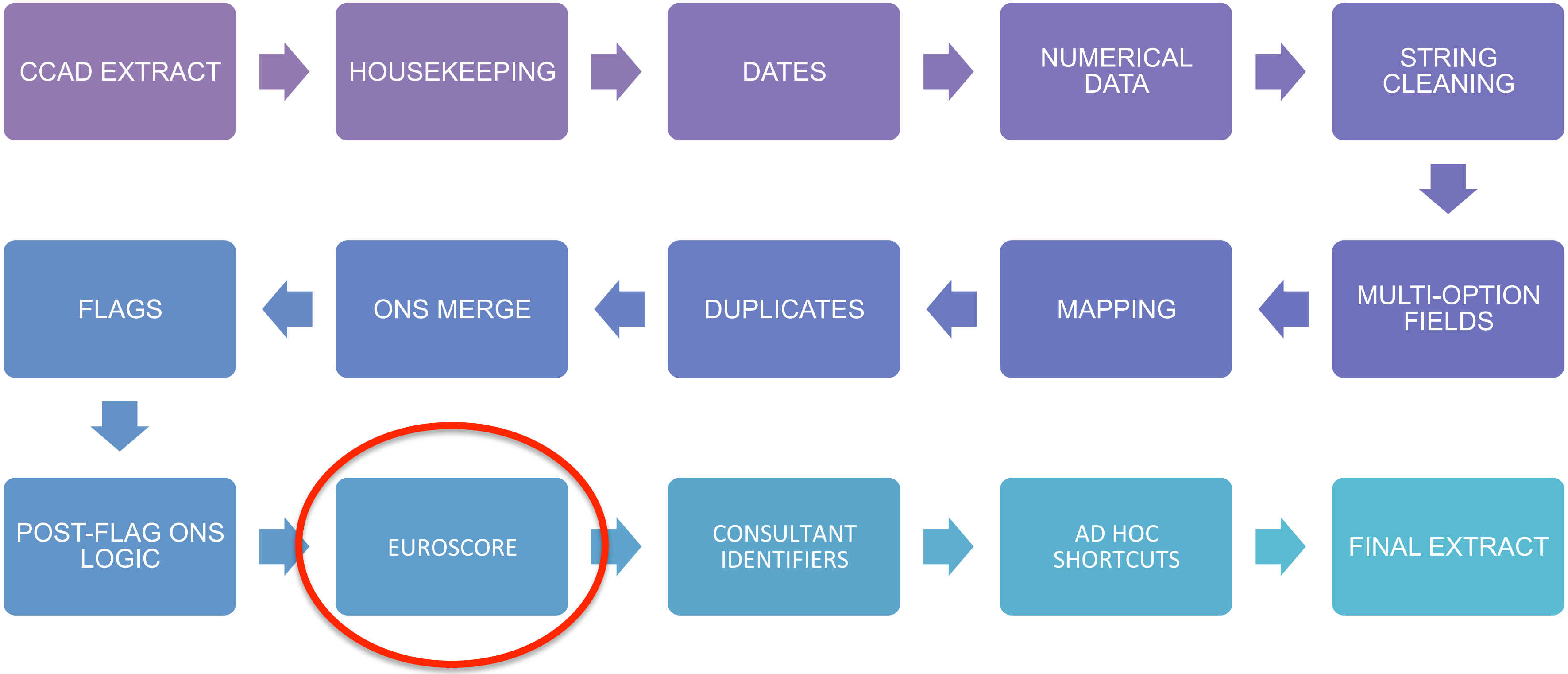
AD HOC SHORTCUTS

FINAL EXTRACT

Flags

- Resolve conflicts
  - in-hospital mortality (e.g. deceased but sent home)
  - back-fill missing mortality from ONS
- Evidence based indicators (incl. resolving conflicts):
  - (individual) valve procedures
  - first operation in a single admission spell
  - first-time cardiac surgery

Cleaning schema

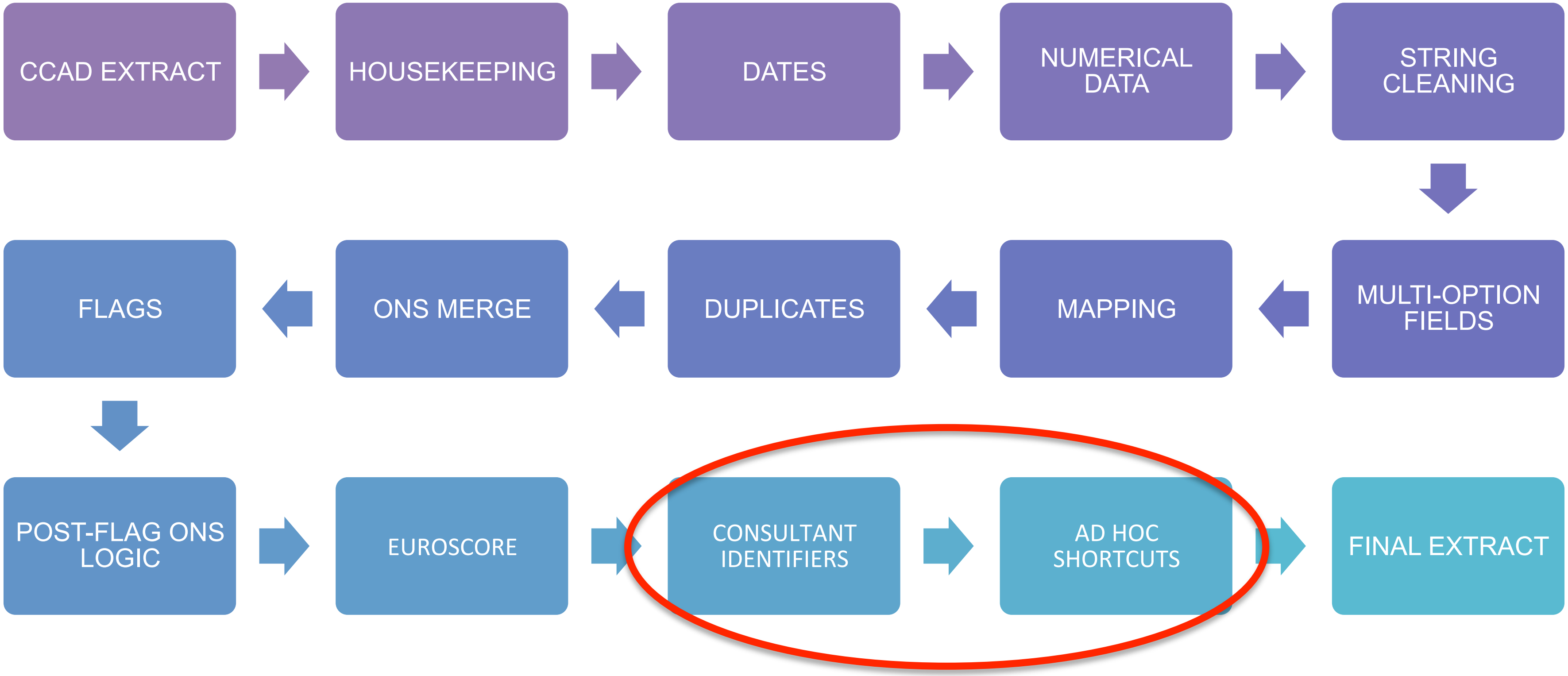


NIBHI

EuroSCORE

- 3 predictions calculated: logistic, mEuroSCORE & EuroSCORE II
- Emphasis on identifying true missing values:
  - data quality measure
  - future analysis of consequences of SCTS imputation
- Database not developed with EuroSCORE II in mind

Cleaning schema



NIBHI

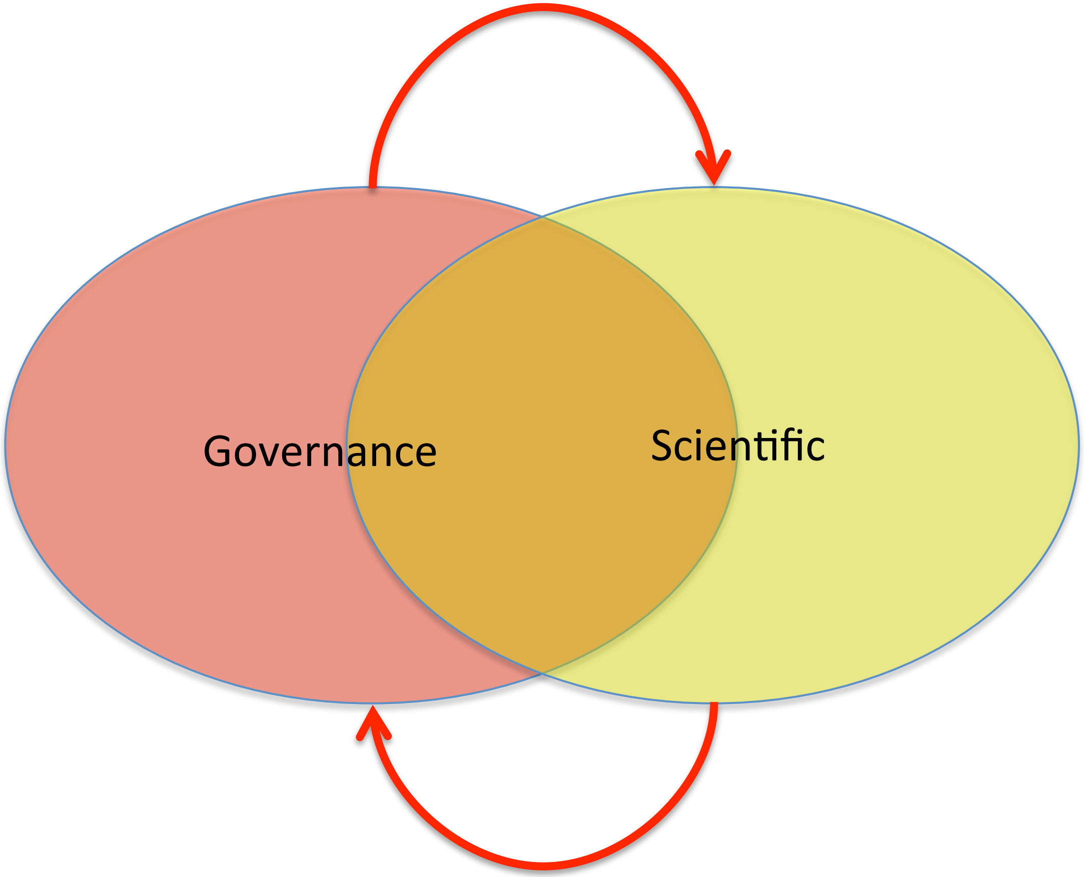
Additional modules

- Consultant identifiers coded to GMC numbers
  - GMC database; hospital webpage; Dr. Forster
- Records deleted for serious ONS date discrepancies
- Expanding list of shortcut fields (e.g. country, financial year)

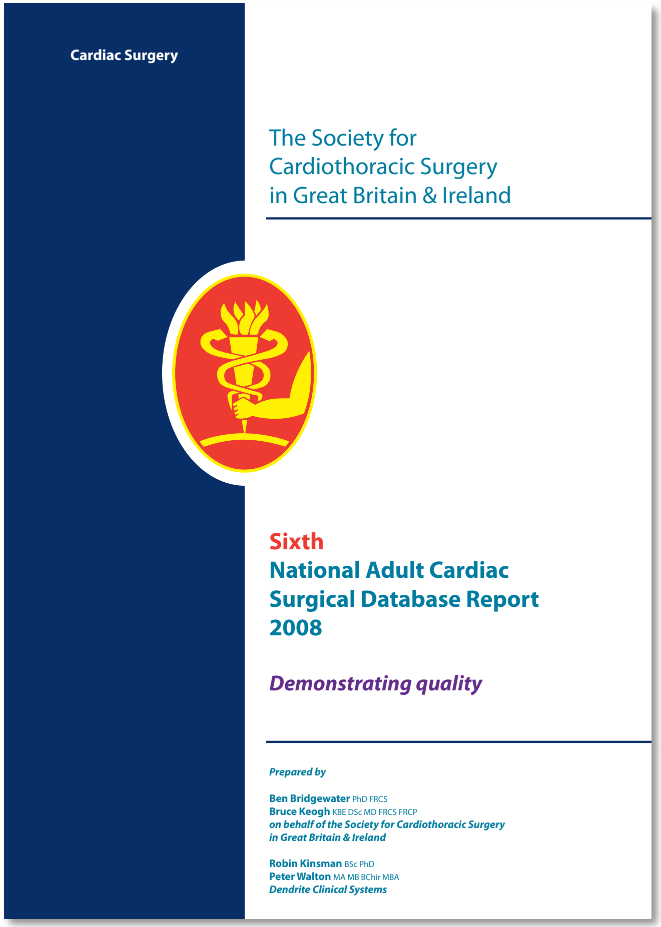
Future cleaning

- Trust-level publication of deleted records
- Tweaks based on validation feedback
- Revisit assumptions + ‘quick-fixes’ of numerical values
- Refinement of the aortic field mappings
- Centralized cleaning / mapping by NICOR

Analyzing the data



Governance



Cardiac Surgery

The Society for Cardiothoracic Surgery in Great Britain & Ireland

Sixth
National Adult Cardiac Surgical Database Report 2008

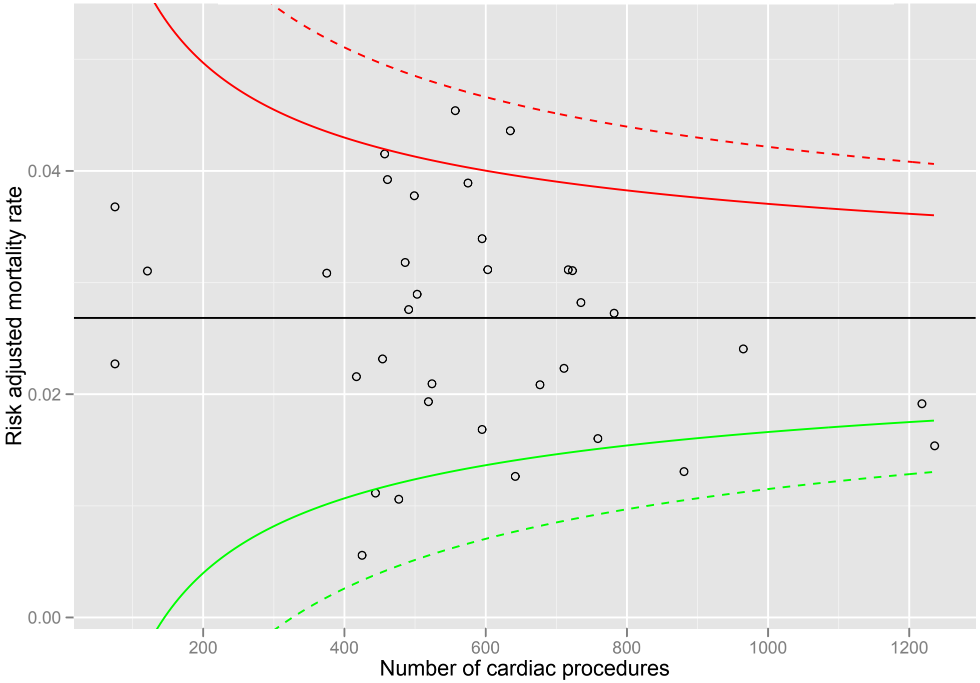
Demonstrating quality

Prepared by

Ben Bridgewater PhD FRCS
Bruce Keogh KBE DSc MD FRCS FRCP
on behalf of the Society for Cardiothoracic Surgery in Great Britain & Ireland

Robin Kinsman BSc PhD
Peter Walton MA MB BChir MBA
Dendrite Clinical Systems

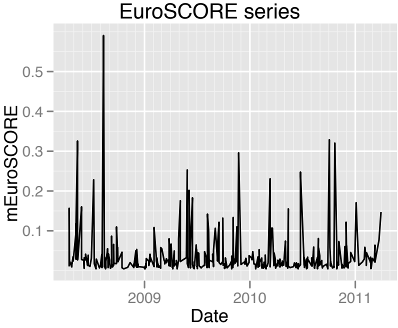
EuroSCORE II: all cardiac surgery

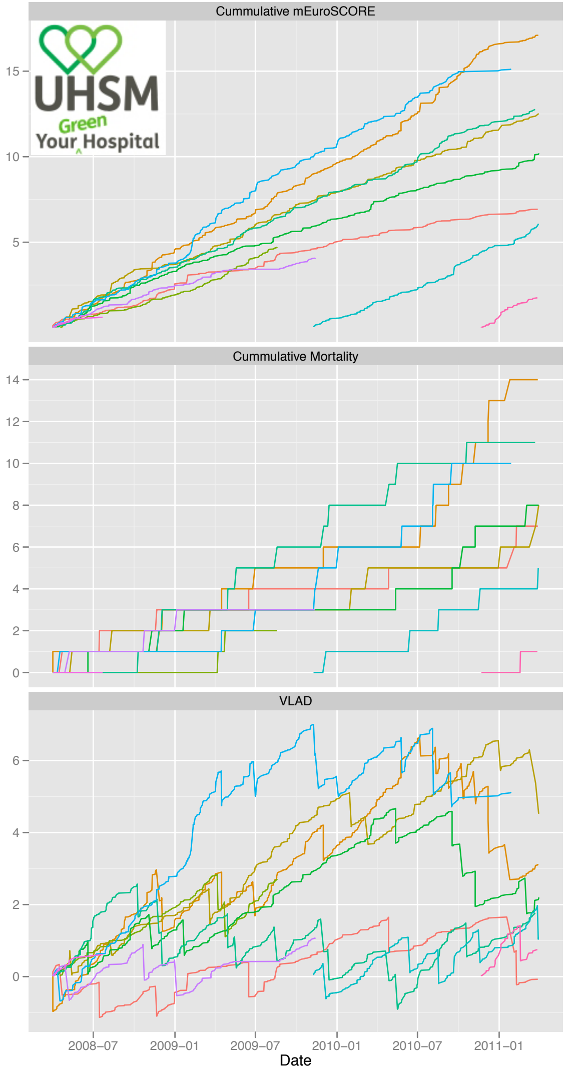


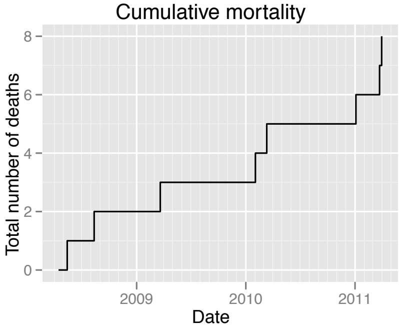
NWEH

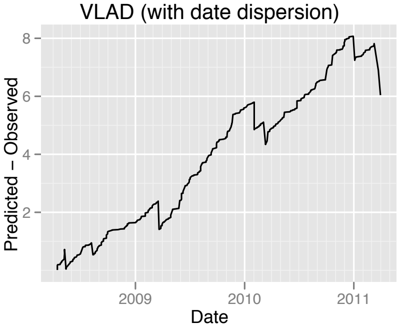
making sense of local healthcare

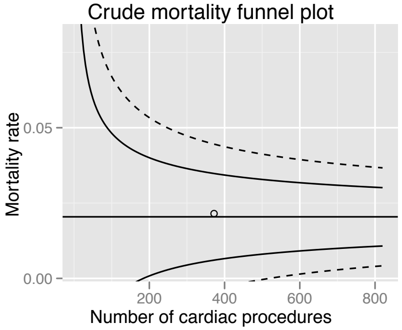
Informing our members

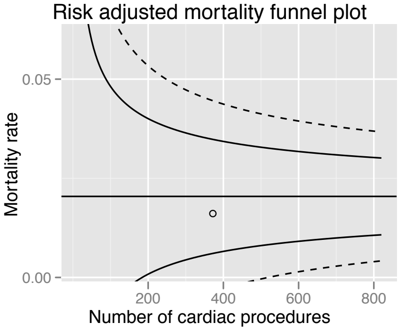




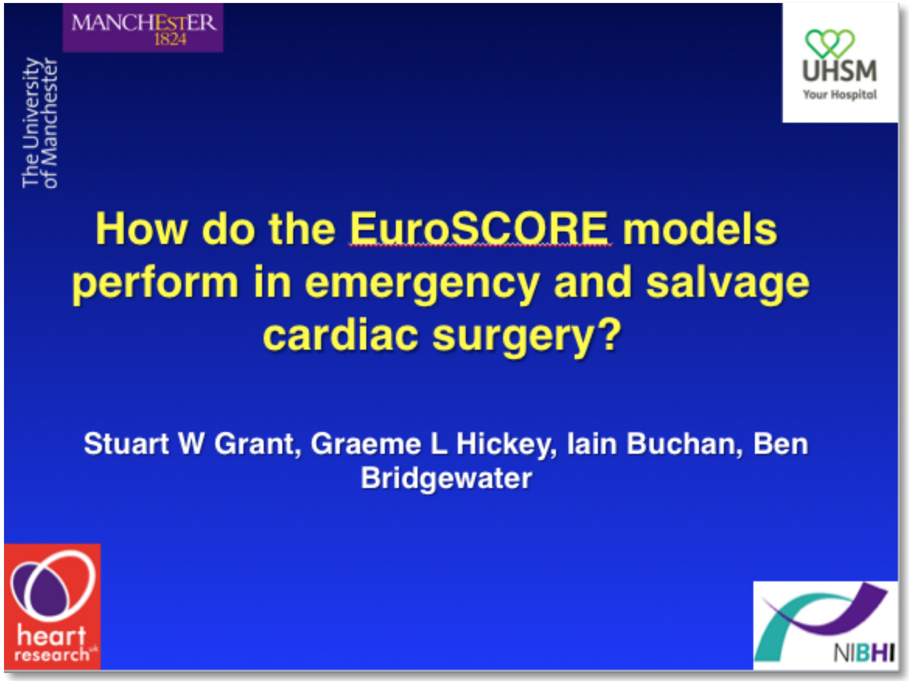






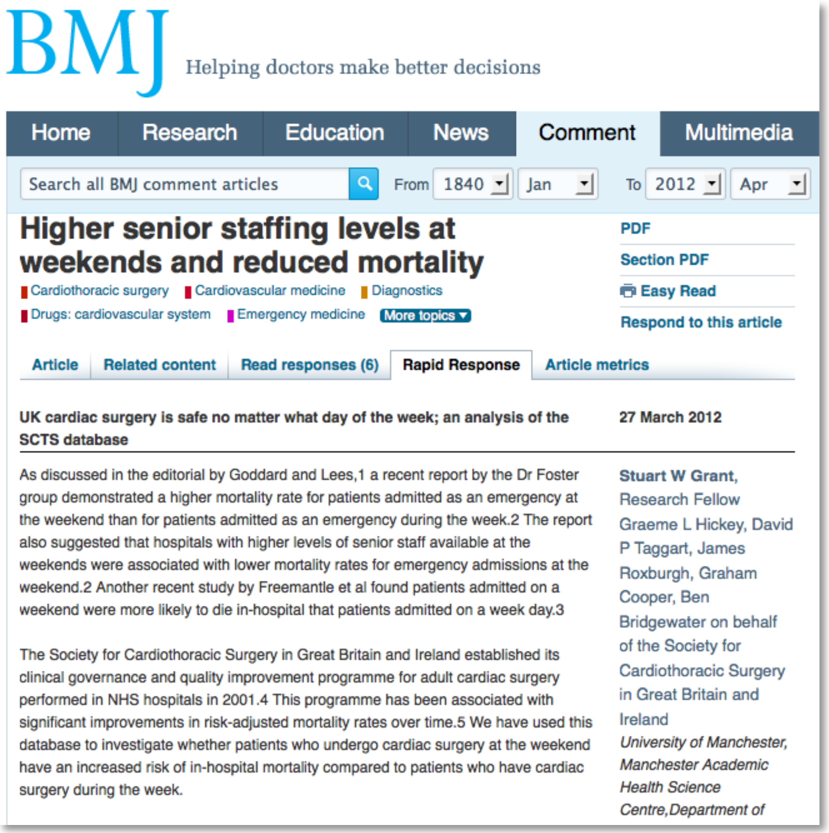


Responding to contemporary questions


MANCHESTER
1824
The University of Manchester
UHSM
Your Hospital
heart research
NIBHI

How do the EuroSCORE models perform in emergency and salvage cardiac surgery?

Stuart W Grant, Graeme L Hickey, Iain Buchan, Ben Bridgewater



Higher senior staffing levels at weekends and reduced mortality

UK cardiac surgery is safe no matter what day of the week; an analysis of the SCTS database

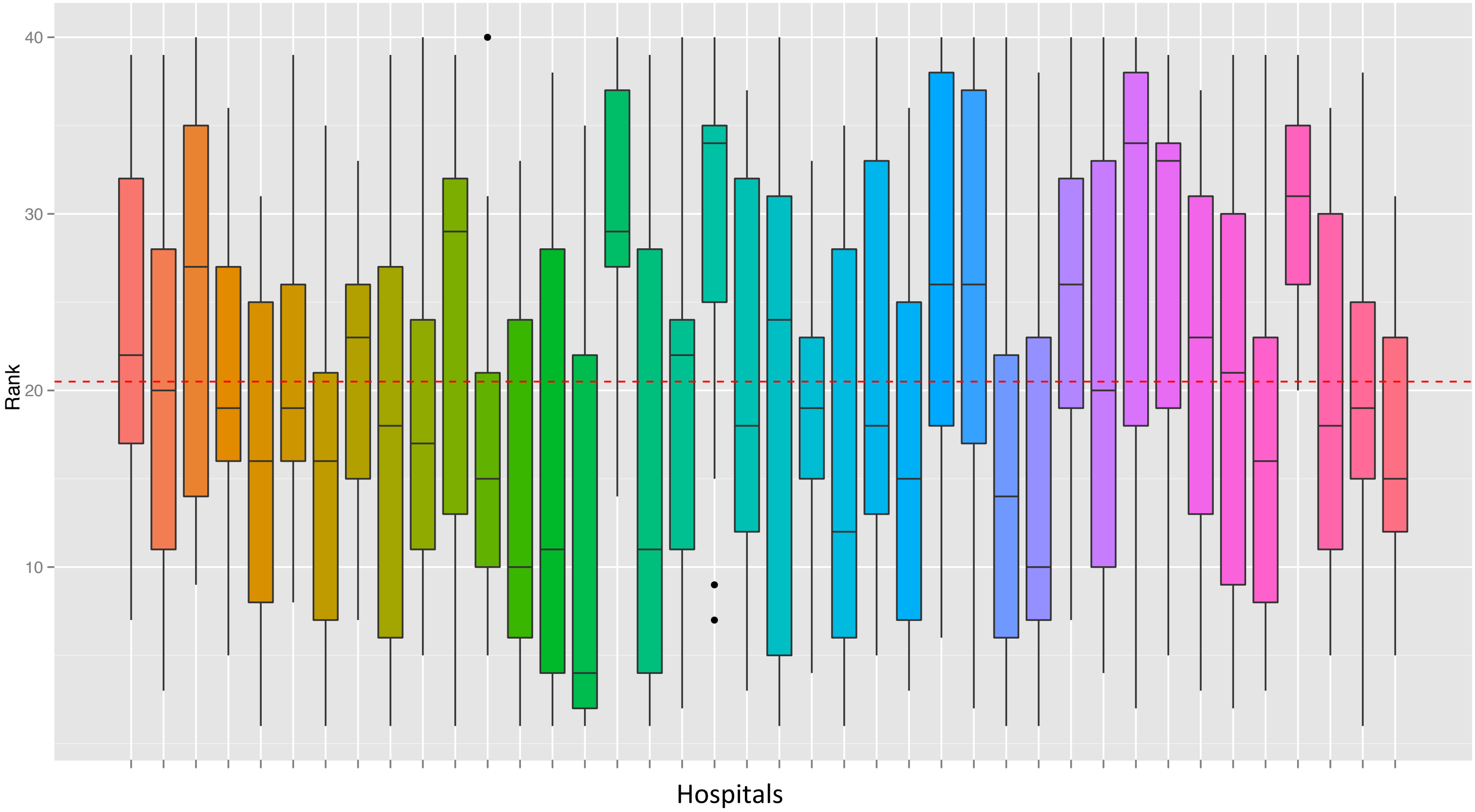
27 March 2012

As discussed in the editorial by Goddard and Lees,1 a recent report by the Dr Foster group demonstrated a higher mortality rate for patients admitted as an emergency at the weekend than for patients admitted as an emergency during the week.2 The report also suggested that hospitals with higher levels of senior staff available at the weekends were associated with lower mortality rates for emergency admissions at the weekend.2 Another recent study by Freemantle et al found patients admitted on a weekend were more likely to die in-hospital that patients admitted on a week day.3

The Society for Cardiothoracic Surgery in Great Britain and Ireland established its clinical governance and quality improvement programme for adult cardiac surgery performed in NHS hospitals in 2001.4 This programme has been associated with significant improvements in risk-adjusted mortality rates over time.5 We have used this database to investigate whether patients who undergo cardiac surgery at the weekend have an increased risk of in-hospital mortality compared to patients who have cardiac surgery during the week.

Stuart W Grant, Research Fellow
Graeme L Hickey, David P Taggart, James Roxburgh, Graham Cooper, Ben Bridgewater on behalf of the Society for Cardiothoracic Surgery in Great Britain and Ireland
University of Manchester, Manchester Academic Health Science Centre,Department of

Measuring data quality



Distribution of ranks of EuroSCORE risk factor prevalence might be expected to homogenous across hospital

↓

Further investigation required

Scientific

- Mitral valve prosthesis: mechanical vs. biological
- Model validation (→ ensure current governance)
- Calibration drift detection methodology (→ inform future governance)

Further information

- SCTS website
  - www.scts.org/
- SCTS-NIBHI project website (incl. contacts)
  - personalpages.manchester.ac.uk/staff/graeme.hickey/scts/
- NICOR website
  - www.ucl.ac.uk/nicor

Acknowledgements

- Heart Research UK – funding
- Sue Manuel (NICOR) – database extracts
- All hospital audit leads and database managers – validating audit summaries
- UK cardiac surgeons – ensuring the validity and accuracy of the data inputted
- The SCTS and all its members – for supporting the audit project