



SCTS Database Cleaning

Version 7.0

Graeme Hickey¹, Stuart Grant², Norman Stein¹ and Ben Bridgewater^{1,2}

¹Northwest Institute for Bio-Health Informatics, Manchester University

²University Hospital of South Manchester, Department of Cardiothoracic Surgery

Document prepared on 21st October 2012

Summary

In this report we provide a brief outline of the steps involved in cleaning the SCTS database with specific details in places where it is deemed necessary to understand the data pre-processing. The database is cleaned using the statistical package R that also comes with base packages in data frame management and regular expression coding. Previously scripts were in a modular format, however recently they have undergone a restructuring to be isolated field scripts.

Data

The database is comprised of a concatenation of two versions of the SCTS databases: V3.8 and V4.1.2. The database is extracted as a single file, although some fields are version-specific. The general strategy of the cleaning process is to merge and map V3.8 into V4.1.2 format in order to allow for straightforward 'overall' cleaning and future analysis. During the transition some hospitals appeared to have completed both database formats, therefore cleaning the databases separately may have led to discarded data. The primary source of information used to establish cleaning and merging routines were found in the database definition documents which are Microsoft Excel spreadsheet files that describe each field including the available options, definitions and format. Both definitions documents (V3.8 and V4.1.2) can be downloaded from here:

V3.8:

<http://www.ic.nhs.uk/webfiles/Services/NCASP/Heart/Datasets/SCTSDataset.xls>

V4.1.2

http://www.ic.nhs.uk/webfiles/Services/NCASP/Heart/Newwebdocuments/SCTS_dataset_ver_4_1_2,_update.xls

Validation

The database has gone through a number of validation exercises both internally and externally. During November 2011, all participating trusts were asked to review their data on the basis of a number of automatic reports generated and a deadline of January 2012 for updating records was set. A repeat of the validation exercise in England and Wales was undertaken in February 2012 (with a March 2012 re-submission date) as part of the National Audit project.



Acknowledgements

We acknowledge all members of the SCTS who have contributed data to the SCTS database and worked with the authors to ensure that scientific reports have been to the highest quality. We are grateful to all database managers working in public and private hospitals who have collected, uploaded and validated data, and those who helped improved the data processing procedure. We thank Kate McAllister and Iain Buchan (Northwest Institute of Bio-Health Informatics) who supported the database-cleaning project through informative discussions.



Table of Contents

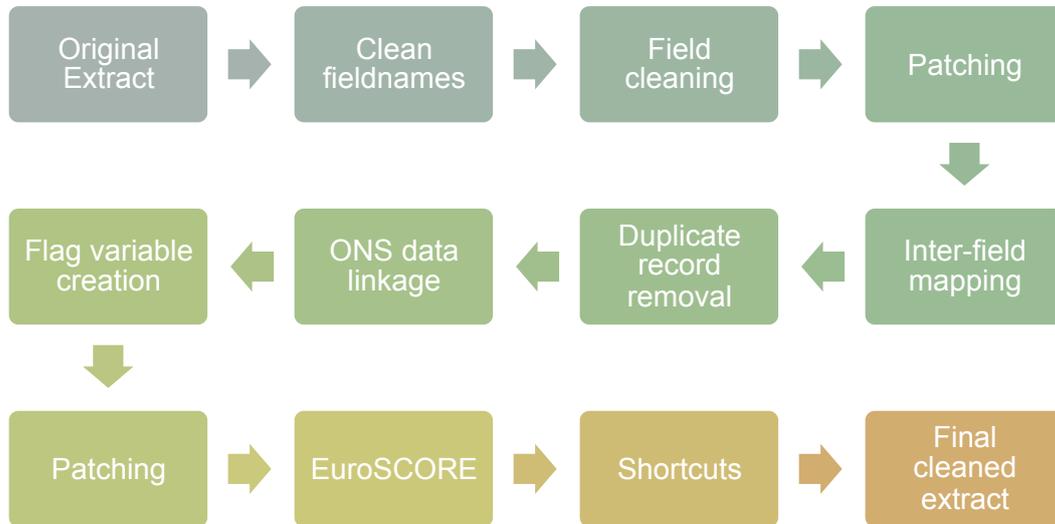
Summary	1
Data	1
Validation	1
Acknowledgements	2
General Schematic of Database Development	5
Points of Record Removal	5
Fieldnames	7
Variable Cleaning	7
<i>Transcriptional String Errors</i>	7
<i>Dates</i>	8
<i>Transcriptional numerical errors</i>	9
2.23 PA Systolic	9
2.24.1 Severity of AVS EOA	10
2.24.2 Severity of AVS Gradient	10
2.25 Left Ventricular End Diastolic Pressure	10
2.26 Mean Pulmonary Artery Wedge Pressure	10
2.27 Ejection Fraction	10
2.36 Number of Previous Heart Operations	10
2.37 Height	10
2.38 Weight	10
3.14 Number of Grafts	11
3.18 Number of Valves Repaired/Replaced	11
3.54 Aortic valve or ring size; 3.58 Mitral valve or ring size; 3.62 Tricuspid valve or ring size &	
3.66 Pulmonary valve or ring size	11
3.67 Number of Aortic Segments operated on	11
3.85 Cumulative Bypass time	11
3.86 Cumulative Cross Clamp Time	11
3.87 Circulatory Arrest Time	12
Age at operation	12
<i>Mapping</i>	12
2.12 Renal	12
2.13 History of Pulmonary Disease	12
2.19 Left Heart Catheterisation	12
2.24 Aortic Valve Gradient	13
2.28 LV Ejection Fraction Category	13
2.33 Intra-aortic Balloon Pump	13
2.34 Reason for Intra-aortic Balloon Pump Used	13
3.13 Other thoracic and vascular procedures	14
Aortic Pathologies	14
Aortic Procedures	15
3.90 Aetiology	16
3.11 Cardiac Procedures	17
3.81 Cardioplegia – Temperature	17
3.82 Cardioplegia – Infusion mode	17
Post-cleaning Inter-field Mapping	17
<i>Aortic Pathologies</i>	17
<i>Procedure Indicators</i>	18



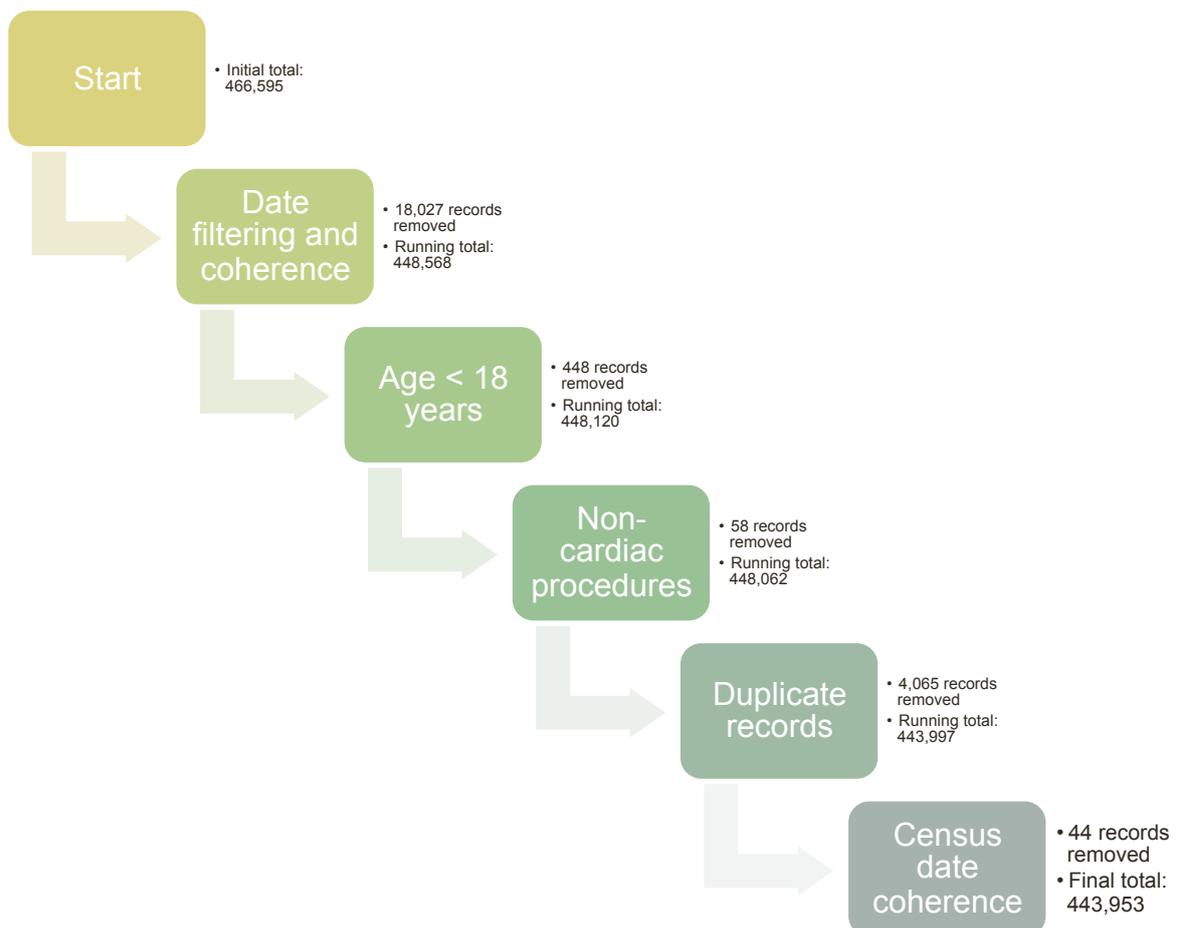
Patching	18
<i>Aortic procedures</i>	19
<i>Valve procedures</i>	19
Removed Records	19
<i>Date Conflicts</i>	19
<i>Adult Status</i>	20
<i>Non-cardiac Procedures</i>	20
<i>Duplicate Records</i>	20
ONS Data Merge	20
Flags	21
LV Ejection Fraction	21
Dead Flag	21
Dead Flag 2	21
Valve Procedures	22
Single Episode	22
Previous Cardiac Operation	23
First Time Cardiac Procedure	23
CABG Flag	23
Valve Flag	24
EuroSCORE	24
Shortcuts	24
Database Issues & Requests	25
<i>High Priority</i>	25
<i>Medium Priority</i>	25
<i>Low Priority</i>	25



General Schematic of Database Development



Points of Record Removal







Fieldnames

The first cleaning script handles some 'house keeping' necessary to initiate the cleaning process.

- Fieldnames are cleaned, shortened and changed where appropriate. For example, some fieldnames have spelling errors in them; some have arbitrary spaces (represented as dots in the R platform); and some fieldnames have changed name from extraction to extraction such as 'Age at Procedure' (in V3.8) to 'Age at Operation' (in V4.1.2).
- Note that some fields previously collected in V3.8 are now considered redundant (e.g. 2.06; 2.12; 2.15; 2.24; 3.11 and 3.13). However, they are required for mapping the existing information to new fields and are therefore retained.
- The following fields are deleted since they were considered to be irrelevant or duplicate data: 'Hospital' (repeated), 3.31, 3.32, 3.33, 3.34, 3.39, 3.40, 3.41, 3.42, 3.51, 3.52, 3.55, 3.56, 3.59, 3.60, 3.63 and 3.64. Also, the following fields were deleted: 'Flow in infarct related artery' (which was empty) and 'Prognostic Score' (generally empty). Further fields are deleted later onwards after the cleaning process.

Variable Cleaning

The errors in the SCTS database have arisen in a number of forms and via different mechanisms. Some cleaning required specialist attention and this is described separately later on. We have broadly categorised errors into the following categories:

1. Transcriptional string errors.
2. Transcriptional numerical errors.
3. Date logic errors.
4. Mapping errors.

In addition to these errors, it is necessary to incorporate mapping between fields and options in V3.8 and those in V4.1.2. Whilst in principle mapping can occur after cleaning, it makes sense in the interest of brevity to combine the cleaning efforts.

Transcriptional String Errors

String options are those that are selected through drop-down menus and radio button controls. In principle they should be error free, but due to historical data, different hospital software, ad hoc record editing and errors at the central repository, errors have inevitably emerged. We generally resolved string errors in a two-step approach:

1. Sweeping manual corrections;
2. Automated self-learning macro corrections.

Manual corrections were generally necessary where the option had been substituted with free text. In cases where free text was entered but with the correction option number, the automated corrector was frequently successful. Manual corrections were embedded in the cleaning scripts retrospectively after monitoring the outputs of the automated scripts.



Reasons for the string cleaning include discrepancies with respect to:

- Case sensitivity.
- Spelling.
- Text file encoding (i.e. the occurrence of symbols).
- Option number contradictions (e.g. '1.' prefixing a record which should have a '2.' prefix).
- Junk inputs (e.g. dates in numerical fields).
- Vital number prefix omission (where it could not be determined automatically).
- Ad hoc mapping.

As the database expands some of the manual correction routines have become redundant. Nonetheless, they are retained for double-robustness in cleaning and also for subsample cleaning.

In some cases where an invalid and indeterminate (i.e. just a number) option had been entered into a field, rather than clean it, per se, a sensible approach taken is to map the option into 'other' or 'unknown' options where they exist. For example, selecting '5' for '3.68.1 Aortic pathology – Ascending Segment Code 2' is invalid; therefore we map option 5 to '99. Other'. The reasoning behind when and when not to take this approach was based on complicated expert judgement.

Automated cleaning was designed to:

- Homogenise inputs (e.g. '1.' and '1. Yes' mapped to a common option '1. Yes').
- Remove options outside of specified bounds (e.g. selection '7. Mitral' when only options 1-4 allowed).
- Handle the allowable 'Unknown' options (usually denoted as '9.' or '99.').
- Choose between multiple inputted valid options for a field that only allows for one value by specifying a *trump mechanism* on a case-by-case basis. This is specified manually.

Multi-option fields were also automatically cleaned. This involved the same principals as above, except additional care was required in handling the separations. One notable difference included where a multi-option was in conflict, two contrasting examples include:

1. '4.02 New post-operative neurological dysfunction': a patient was recorded as having '1. Transient stroke' and '2. Permanent stroke'. In this case option 2 trumped option 1.
2. '3.12 Other Actual Cardiac Procedures': the response '0. No other cardiac procedures performed' was included¹ with an actual other cardiac procedure. In this case the field was in conflict and the input set as missing.

Remaining string errors are handled by manual post-correction. No specific examples of variables that are string cleaned are provided here; please consult the cleaning code for this.

Dates

Dates are stored in the SCTS database as date-time strings. However, the time strings for many fields are default, e.g. 00:00 or 00:01. Therefore we delete all time strings with the exception of the

¹ The SCTS V4.1.2 database description document does not include the option '0. No other cardiac procedures performed' nor any mention of it in the 'fields to remove' or 'fields to change' sections. We therefore retained this option due to it being helpful in populating field 3.11.4 during mapping.



procedure date times (which are reasonably well recorded) that will allow us to chronologically order procedures for determining the first cardiac procedure in any admission spell.

For the following fields, all dates after 20/03/2012 (the date of the database extraction) were set as missing:

- Date
- Creation Date
- 3.01 Admission Date
- 3.02 Procedure Date
- 4.06 Discharge Date

For the following fields, all dates not in the range 01/01/1967 (a proxy date for cardiac surgery in the UK) – 20/03/2011 were set as missing:

- 2.08 Date Last Cardiac Operation
- 2.20 Date Last Cardiac Catheterisation

Transcriptional numerical errors

Numerical inputs are those that require just a number, e.g. height and weight. Adjustments are made to avoid implausible values that may have been entered accidentally; due to misunderstanding; or because of conflicting individual hospital central data management.

For a handful of variables that are not routinely used in governance or scientific research (e.g. valve sizes) we have only minimally cleaned them. This would need to be reviewed later on if the data was to be used. Often the primary reason for a numerical error is that the field was set as a free-text field in the input software thus allowing any alphanumeric entry.

In cases where text or symbols had been entered into the variable, they were automatically removed. In some cases this meant removing symbols used to indicate bounds or ranges, e.g. <, >, -. Only for variables known to be required for calculation of risk scores was special care taken in the cleaning of these boundaries, albeit with ad hoc rules. In general the number of these cases was small (<0.01%) and therefore irrelevant.

We not list all transcriptional numerical error corrections. Those listed below give details of specific assumptions made by the authors of the cleaning software.

2.23 PA Systolic

Pressures not inside the interval (0, 200] mmHg were set as missing. There were > 330,000 records recorded as 0mmHg in addition to other unlikely values, e.g. 10mmHg. It is known that surgeons indicate 'normal pressure' using 0 mmHg, hence the values are subsequently retained for later risk factor analyses. These pressures should therefore **not** be considered as actual values.

Previously, all values for Papworth Hospital were defined as missing since they only record 61 mmHg as code for > 60 mmHg or 'high'. We did not invoke this rule because, despite being improperly recorded, the indicator of < or > 60 mmHg is still important for the [modified] EuroSCORE model.

All data for Morriston Hospital prior to 01/04/2006 is defined as missing due to an apparent incorrect specification.



2.24.1 Severity of AVS EOA

Any values not inside the interval [0.1, 6] were set as missing.

2.24.2 Severity of AVS Gradient

Any values not inside the interval [15, 200] were set as missing.

2.25 Left Ventricular End Diastolic Pressure

Any values not inside the interval (0, 300] were set as missing.

2.26 Mean Pulmonary Artery Wedge Pressure

All values not inside the interval (0, 100) were set as missing.

2.27 Ejection Fraction

Where a minimum or maximum threshold was given (i.e. > or <) the numerical values were adjusted by +1 and -1 respectively.

All negative values were ignored and absolute values taken.

All values not inside the interval [15, 100) were set as missing.

2.36 Number of Previous Heart Operations

Only 0-6 previous heart operations allowed; values not inside the interval [0, 6] are defined as missing.

2.37 Height

A sequence of rules were used to clean the heights based on a number of observations:

1. Values recorded as exactly '2.000' were set as missing.
2. Values inside the interval (1.4, 2.2) were multiplied by 100 based on the assumption they were originally recorded in meters and not centimetres; this was confirmed via clustering.
3. Values inside the interval (12,000, 22,000) were divided by 100 based on the assumption they were recorded in millimetres and not centimetres.
4. Remaining values not inside the interval [107, 250] were set as missing.

2.38 Weight

Values not inside the interval [25, 250] kg were set as missing.

3.03 Responsible Consultant Surgeon

Consultant identifiers were recorded using a variety of methods, including: GMC numbers (including a variety of prefixes); names (given name and/or surname and/or title) and initials. An algorithm was written which implements the following steps:



1. Junk inputs were removed automatically.
2. Prefixes were stripped from GMC numbers.
3. Where 2 or more consultants were listed, only the first one was retained.
4. Manual conversion of names and initials to GMC numbers. Sources for matching surgeons to GMC numbers are:
 - a. Master surgeon document compiled by Mr Ben Bridgewater;
 - b. Dr Foster Health (<http://www.drfoosterhealth.co.uk>);
 - c. Royal College of Surgeons (<http://www.rcseng.ac.uk/>);
 - d. Society of Cardiothoracic Surgery (<http://www.scts.org>);
 - e. GMC (<http://www.gmc-uk.org/>).
5. GMC numbers for surgeons practicing at Mater Misericordiae University Hospital could not be determined (although multiple consultants were still reduced to the first recorded).
6. A few surgeons could not be identified or their GMC number found, e.g. due to surgeon being retired/deceased. These were left as surnames or initials, whatever was originally inputted.
7. A small number of records had non-cardiothoracic consultants listed (e.g. vascular consultants or SpRs). These were only ever associated with one or two records so the consultant identifier was set to missing.

3.14 Number of Grafts

Integer values outside the interval [0, 11] were set as missing.

3.18 Number of Valves Repaired/Replaced

Integer values outside the interval [0, 4] were set as missing.

3.54 Aortic valve or ring size; 3.58 Mitral valve or ring size; 3.62 Tricuspid valve or ring size & 3.66 Pulmonary valve or ring size

Values outside the interval [10, 50] mm² were set as missing.

3.67 Number of Aortic Segments operated on

Integer values outside the interval [0, 5] are defined as missing.

3.85 Cumulative Bypass time

Symbols were retained and where appropriate the arithmetic sums evaluated to yield a single value.

Remaining values not inside the interval [0, 10,080] minutes were set as missing.

3.86 Cumulative Cross Clamp Time

Values not inside the interval [0, 360] minutes were set as missing.



3.87 Circulatory Arrest Time

Values not inside the interval [0, 240] minutes were set as missing.

Age at operation

Values that were negative by > 5 years are incremented by 100 years based on the assumption that the hospital database software inputted the wrong century.

Remaining values not inside the interval [0, 99) were set as missing. This includes those in the interval [-5, 0).

Mapping

The most problematic element of the database restructuring is the issue of mapping. This arises due to: 1) SCTS V4.1.2 includes new fields which requires inter-field mappings from SCTS V3.8; and 2) SCTS V4.1.2 has redefined field options which requires intra-field mappings from the existing SCTS V3.8. Another issue is that the database definitions document does not provide adequate mappings for all field changes. For example, '2.29 Intravenous nitrates or any heparin' no longer allows for the option '2. Within one week', but no decision is specified on whether to map this to options '0. No' or '1. Yes'.

For some fields we opted not to map between the database structures. For example, options 6-9 were dropped from '2.07 Previous cardiac surgery' in the switch from SCTS V3.8 to SCTS V4.1.2, but rather than delete the old options, they were retained. Here we describe all mappings that implemented **concomitantly** with the cleaning.

2.12 Renal

Inter-field mapping of V3.8 field 2.12 to V4.1.2 fields 2.12.0 and 2.12.1 based on clinical judgments. This included a pseudo-mapping of acute and chronic creatinine levels to 80 and 250 $\mu\text{mol/l}$; see table below.

SCTS V3.8 2.12	SCTS V4.1.2 2.12.0 ($\mu\text{mol/l}$)	SCTS V4.1.2 2.12.1
0. No renal disease	80	0. None
1. Functioning transplant	80	0. None
2. Creatinine > 200 $\mu\text{mol/l}$	250	Missing
3. Dialysis for acute renal failure: onset within 6 weeks of cardiac surgery	250	1. Dialysis for acute renal failure: onset within 6 weeks of cardiac surgery
4. Dialysis for chronic renal failure: onset more than 6 weeks prior to cardiac surgery	250	2. Dialysis for chronic renal failure: onset more than 6 weeks prior to cardiac surgery
9. Unknown	Missing	Missing

The old 2.12 field and an indicator of whether 2.12.0 was originally recorded or missing (prior to pseudo-value imputation) are added to the database to support sensitivity analyses.

2.13 History of Pulmonary Disease

Intra-field mapping of V3.8 field options '1. COAD/emphysema' and '2. Asthma' to new V4.1.2 field option '1. COAD/emphysema or Asthma'.

2.19 Left Heart Catheterisation



Intra-field mapping of V3.8 field option '2. Previous admission' to V4.1.2 field option '0. No'.

2.24 Aortic Valve Gradient

Missing values in '2.24.2 Severity of AVS Gradient' were replaced by the corresponding values in 2.24 (which may also be missing). Note that all data for Papworth in 2.24 were defined as missing.

2.28 LV Ejection Fraction Category

Intra-field mapping of '1. Good (LVEF \geq 50%)' to '1. Good (LVEF > 50%)'. A large number of records with the old '1. Good (LVEF \geq 50%)' category had a measured ejection fraction in 2.27 of precisely 50%. We assume that 50% was entered as an indicator of good LVEF and therefore we **do not** map these records into '2. Fair (LVEF 30-50%)' category. However, where the category (2.28) was missing and a measured ejection fraction (2.27) was recorded as 50%, we set 2.28 category to '2. Fair (LVEF 30-50%)'; this affected very few cases.

2.33 Intra-aortic Balloon Pump

Inter-field mapping of V3.8 fields '2.33 Intra-aortic balloon pump used' (a multi-option field) to 12 new fields in V4.1.2. A suffix of '.0', '.1', '.2' or '.3' indicates whether an IABP, Impeller, Ventricular Assist or Other device was used respectively. A *further* suffix of '.1', '.2' or '.3' indicates whether the device was used pre-op, intra-op or post-op. Hence $4 \times 3 = 12$ fields.

The database definitions document describes the addition of only 4 fields per 2.33 and 2.34 respectively; i.e. excluding the pre-op, intra-op and post-op indicators respectively. However, the included in the document is an appended 'corrections sheet' which confirms the inclusion on the surgical-stage aspect (as marked by the final suffix) as a new addition. Nevertheless, the field options are contradictory. '2.33.0.1 IABP Preop', '2.33.0.2 IntraOp' and '2.33.0.3 PostOp' each allow for one of 4 options: 1. No; 2. Pre-operation; 3. Intra-operation; 4. Post-operation. These options conflict with the field definitions. Furthermore, surgeons have actually chosen contradictory options. It is conjectured that SCTS were meant to have set these field options to '1. Yes' or '0. No'.

The following action is taken:

- 4 new multi-option fields are created by collapsing the surgical-stage suffixes:
 - '2.33.0 Intra-aortic balloon pump used'
 - '2.33.1 Impeller device used'
 - '2.33.2 Ventricular assist device used'
 - '2.33.3 Other Support Device used'
- The following fields are deleted: 2.33.0.1; 2.33.0.2; 2.33.0.3; 2.33.1.1; 2.33.1.2; 2.33.1.3; 2.33.2.1; 2.33.2.2; 2.33.2.3; 2.33.3.1; 2.33.3.2 and 2.33.3.3.

2.34 Reason for Intra-aortic Balloon Pump Used

Inter-field mapping of '2.34 Reason for intra-aortic balloon pump used' (a single-option field) to 12 new fields in SCTS V4.1.2. A suffix of '.0', '.1', '.2' or '.3' indicates whether the reason corresponds to use of an IABP, Impeller, Ventricular Assist or Other device. A *further* suffix of '.1', '.2' or '.3' indicates whether the device was used pre-op, intra-op or post-op. Hence $4 \times 3 = 12$ fields.

The 2.34 fields which are the reasons for using the 4 device options at the 3 surgical-stages (pre, intra and post) are incompatible with the original SCTS V3.8 fields since we do not know when the device was used. Hence a mapping of 2.34 to 2.34.0.1, 2.34.0.2 and 2.34.0.3 is impossible without access to medical records.

The following action is taken:



- 4 new single-option fields were created by collapsing the surgical-stage suffixes, taking the first reason provided as the option. Providing these fields are not pivotal then this should be a suitable patch. The new fields are:
 - '2.34.0 IABP Ind'
 - '2.34.1 Impeller Ind'
 - '2.34.2 Ventricular Ind'
 - '2.34.3 Other Ind'
- The following fields are deleted: 2.34.0.1; 2.34.0.2; 2.34.0.3; 2.34.1.1; 2.34.1.2; 2.34.1.3; 2.34.2.1; 2.34.2.2; 2.34.2.3; 2.34.3.1; 2.34.3.2 and 2.34.3.3.

3.13 Other thoracic and vascular procedures

Inter-field mapping of '3.13 Other thoracic and vascular procedures' (a multi-option field) to '3.12 Other Cardiac Procedures' (a multi-option field).

Mapping rules²:

- '1. Carotid endarterectomy' → '17. Carotid endarterectomy'.
- '3. Other thoracic' → '19. Other procedures not listed above'.
- If 3.12 was originally missing or recorded as '0. No other cardiac procedure performed' and a mapping occurred, the original values were overwritten.

Valve Procedures

No actual mapping of the expired '2. Repair' field to either of the new options: '2. Repair with ring' or '4. Repair without ring' was implemented as it is not always possible to distinguish whether a ring was used or not. This affects fields: '3.43 Aortic valve procedure', '3.44 Mitral valve procedure', '3.45 Tricuspid valve procedure' and '3.46 Pulmonary valve procedure'.

Aortic Pathologies

Inter-field mapping of V3.8 aortic pathology fields, '3.68 Aortic pathology – Root', '3.70 Aortic pathology – Ascending', '3.72 Aortic pathology – Arch', '3.74 Aortic pathology – Descending' and '3.76 Aortic pathology – Abdominal' to V4.1.2 aortic pathology fields, '3.68.1 Aortic pathology - Root Segment Code 1', '3.70.1 Aortic pathology - Ascending Segment Code 2', '3.72.1 Aortic pathology - Arch Segment Code 3', '3.74.1 Aortic pathology - Descending Aorta Segment Code 4' and '3.76.1 Aortic pathology - Abdominal Segment Code 5'.

Mapping rules for pathology field (applies to all 5 pathology segments):

- '1. Aneurysm' → '1. Aneurysm'
- '2. Syphilis' → '1. Aneurysm'
- '3. Dissection' → '2. Chronic Dissection' if 2.35 Operative Urgency not recorded as '1. Elective'
- '3. Dissection' → '3. Acute Dissection' if 2.35 Operative Urgency recorded as '1. Elective'
- '3. Dissection' → '99. Other' if 2.35 Operative Urgency missing
- '4. Transection' → '4. Trauma'
- '5. Coarctation' → '99. Other'
- '6. Atheromatous' → '99. Other'
- '7. Marfan's' → '99. Other'

² There is ambiguity over the current interpretation of the current '3.12 Other Cardiac Procedures' field since it now envelopes non-cardiac procedures, e.g. peripheral vascular, pulmonary transplant and carotid endarterectomies. We adopt the SCTS classification system under the presumption the field represents other cardio- pulmonary and vascular procedures.



- '9. Mycotic' → '99. Other'
- '10. Other connective tissue disorder' → '99. Other'
- '11. Congenital' → '99. Other'
- '12. Infection - native' → '99. Other'
- '13. Infection - graft' → '99. Other'
- '99. Unknown' → '99. Other'

In all cases: only empty records in V4.1.2 fields are mapped to; if the record already contains information then no mapping occurs to prevent erasing data from hospitals (very few of them) who appear to have done their own mapping.

Aortic Procedures

Inter-field mapping of V3.8 aortic procedure fields, '3.69 Aortic procedure – Root', '3.71 Aortic procedure – Ascending', '3.73 Aortic procedure – Arch', '3.75 Aortic procedure – Descending' and '3.77 Aortic procedure – Abdominal' to V4.1.2 aortic procedure fields, '3.69.1 Aortic procedure – Root Segment Code 1', '3.71.1 Aortic procedure – Ascending Segment Code 2', '3.73.1 Aortic procedure – Arch Segment Code 3', '3.75.1 Aortic procedure - Descending Aorta Segment Code 4' and '3.77.1 Aortic procedure - Abdominal Segment Code 5'.

Mapping rules:

- Root
 - '3. Root replacement with composite valve graft and coronary reimplantation' → '4. Root replacement with composite valve graft and coronary reimplantation (Modified Bentall or Cabroll)'
 - '4. Root replacement with preservation of native valve and coronary reimplantation' → '5. Root replacement with preservation of native valve and coronary reimplantation'
 - '5. Homograft root replacement' → '6. Homograft root replacement'
 - '6. Autograft root replacement (Ross Procedure)' → '7. Ross Procedure'
 - '7. Aortic patch graft' → '8. Aortic patch graft'
 - '8. Sinus of Valsalva repair' → '9. Sinus of Valsalva repair'
 - '9. Reduction aortoplasty' → '10. Reduction aortoplasty' in the field '3.71.1 Aortic procedure – Ascending Segment Code 2'.
- Ascending
 - '1. Interposition tube graft' → '1. Interposition tube graft with/without extension into the arch'
 - '3. Root replacement with composite valve graft and coronary reimplantation' → '4. Root replacement with composite valve graft and coronary reimplantation (Modified Bentall or Cabroll)'
 - '4. Root replacement with preservation of native valve and coronary reimplantation' → '5. Root replacement with preservation of native valve and coronary reimplantation'
 - '5. Homograft root replacement' → '6. Homograft root replacement'
 - '6. Autograft root replacement (Ross Procedure)' → '7. Ross Procedure'
 - '7. Aortic patch graft' → '8. Aortic patch graft'
 - '8. Sinus of Valsalva repair' → missing
 - '9. Reduction aortoplasty' → '10. Reduction aortoplasty'
- Arch
 - '1. Interposition tube graft' → '2. Interposition tube graft with reimplantation of major vessels'



- '2. Tube graft + separate AVR' → missing
- '3. Root replacement with composite valve graft and coronary reimplantation' → missing
- '4. Root replacement with preservation of native valve and coronary reimplantation' → missing
- '5. Homograft root replacement' → missing
- '7. Aortic patch graft' → '8. Aortic patch graft'
- '9. Reduction aortoplasty' → '10. Reduction aortoplasty'
- Descending
 - '1. Interposition tube graft' → '1. Interposition tube graft'
 - '2. Tube graft + separate AVR' → missing
 - '3. Root replacement with composite valve graft and coronary reimplantation' → missing
 - '4. Root replacement with preservation of native valve and coronary reimplantation' → missing
 - '5. Homograft root replacement' → missing
 - '7. Aortic patch graft' → '8. Aortic patch graft'
 - '8. Sinus of Valsalva repair' → missing
 - '9. Reduction aortoplasty' → missing
- Abdominal
 - '1. Interposition tube graft' → '1. Interposition tube graft'
 - '3. Root replacement with composite valve graft and coronary reimplantation' → missing
 - '4. Root replacement with preservation of native valve and coronary reimplantation' → missing
 - '7. Aortic patch graft' → missing
 - '9. Reduction aortoplasty' → missing

Only empty records in V4.1.2 fields are mapped to: if the record already contains information then no mapping is allowed to prevent erasing data from the small number of hospitals who have retrospectively updated the SCTS database.

3.90 Aetiology

Intra-field mapping of fields 3.68, 3.70, 3.72, 3.74 and 3.76 to 3.90 (multi-option field). Multiple options deriving from either a single or multiple pathology segments are concatenated into a single record which is automatically cleaned afterwards. Information is also sourced from other related fields.

Mapping rules:

- '6. Atheromatous' → '2. Atherosclerosis'
- '7. Marfan's' → '3. The Marfan Syndrome'
- '4. Transection' → '6. Trauma'
- '2. Syphilis' → '10. Aortitis'
- '9. Mycotic' → '9. Infection'
- '12. Infection - native' → '9. Infection'
- '13. Infection - graft' → '9. Infection'
- '10. Other connective tissue disorder' → '10. Aortitis'
- '5. Coarctation' → '7. Coarctation'
- '11. Congenital' → '8. Other congenital'



Mapping rules (using fields other than major aorta pathology data):

- If '2.10 History of Hypertension' is recorded as '1. Treated or BP>140/90 on >1 occasion prior to admission' → '1. Hypertension'
- Use V3.8 options for 2.07.Previous.Surgical.Interventions '5. Aortic surgery - ascending or arch' or '6. Aortic surgery - descending or abdominal' → '11. Previous aortic surgery'.

3.11 Cardiac Procedures

Inter-field mapping of V3.8 field 3.11 'Cardiac Procedures' to V4.1.2 fields '3.11.1 CABG', '3.11.2 Valve', '3.11.3 Major aortic' and '3.11.4 Cardiac Procedures Other'. The mapping rules are as per the table below.

SCTS V3.8 3.11	SCTS V4.1.2 3.11.1	SCTS V4.1.2 3.11.2	SCTS V4.1.2 3.11.3	SCTS V4.1.2 3.11.4
1. CABG alone	1. Yes	0. No	0. No	0. No
2. CABG + valve	1. Yes	1. Yes	0. No	0. No
3. CABG + valve + other	1. Yes	1. Yes	0. No	1. Yes
4. CABG + other	1. Yes	0. No	0. No	1. Yes
5. Valve alone	0. No	1. Yes	0. No	0. No
6. Valve + other	0. No	1. Yes	0. No	1. Yes
8. Other	0. No	0. No	0. No	1. Yes

3.81 Cardioplegia – Temperature

Option '8. Not applicable' mapped to missing. This does not affect records where '8. Not applicable' was with another option, e.g. '1. Warm'. In this case '8. Not applicable; 1. Warm' would map to '1. Warm'.

3.82 Cardioplegia – Infusion mode

Option '8. Not applicable' mapped to missing. This does not affect records where '8. Not applicable' was with another option, e.g. '1. Antegrade'. In this case '8. Not applicable; 1. Antegrade' would map to '1. Antegrade'.

Post-cleaning Inter-field Mapping

The vast amount of mapping rules could be embedded into the individual cleaning code files. This was possible for two reasons:

1. The mapping was an intra-field mapping (i.e. was isolated to the specific field being cleaned).
2. The mapping was an inter-field mapping (i.e. depended on other fields in the database) which had already been cleaned and mapped prior to the current one in the sequence of cleaning.

In two important cases it was not possible to embed the mapping into the cleaning code.

Aortic Pathologies



Where a root segment procedure is listed (options 4, 5 or 6) in the ascending segment procedure field (3.71.1; see database description document for further details) and the root segment procedure field (3.69.1) is empty, a duplicate copy is mapped from the former to the latter. This is to emphasize the root procedure was one which extended into the ascending segment.

Procedure Indicators

The field '3.11.3 Major aortic procedure' is new to V4.1.2. Previously major aortic procedures came under the umbrella of 'other cardiac procedures' in V3.8. Therefore 3.11.3 was not completed prior to 2010 nor were we capable of mapping it from V3.8 field 3.11. Therefore, after cleaning:

- If the fields 3.69.1, 3.70.1, 3.71.1, 3.72.1, 3.73.1, 3.74.1, 3.75.1, 3.76.1 or 3.77.1 were recorded with any legal option, 3.11.3 was set to '1. Yes'.
- If '3.67 Number of Aortic Segments operated on' was > 0 and '3.13 Other thoracic and vascular procedures' was recorded as either '1. Aortic or peripheral vascular' or missing, then 3.11.3 was set to '1. Yes'.

Since 'other cardiac procedures' previously included major aortic procedures in the SCTS V3.8 definition, we needed to separate out these two procedure types (where appropriate) by setting 3.11.4 to '0. No' if both of the following conditions are true:

- 3.11.3 and 3.11.4 were recorded as '1. Yes'.
- 3.12 was recorded as '0. No other cardiac procedure performed' or missing.

Note that this will override some of the initial mappings made for the 3.11 → 3.11.x field series.

Although the field 3.11.4 is new, an error at the central database repository means that we cannot extract this field. Therefore we need to backfill this field (primarily for records submitted to CCAD in V4.1.2 format; though this correction was made for all records [incl. pre-SCTS V4.1.2]) in using evidence of an actual other cardiac³ procedure. Following the other cleaning, mapping and inter-field mapping we backfill using the following rules:

- 3.11.4 was set to '1. Yes' if 3.11.4 is initially missing and 3.12 is not missing or recorded as '0. No other cardiac procedure performed'.
- 3.11.4 was set to '0. No' if 3.11.4 is missing and both of the following conditions are true:
 - 3.12 is missing or recorded as '0. No other cardiac procedure performed'; and
 - 3.13⁴ is missing or recorded as '0. No thoracic and vascular procedures performed'.

Patching

Obvious database record errors that were isolated to one particular hospital and /or one particular time period were 'patched' pending a resolution at the trust-level or central database repository. In

³ 'Other cardiac procedures' actually includes other cardiothoracic and vascular procedures.

⁴ 3.13 is leveraged here because the historical option of '1. Aortic or peripheral vascular' does not neatly map into the V4.1.2 format since an aortic procedure would map into 3.11.3 and a peripheral vascular would map into 3.12 and 3.11.4.



these cases, trusts were contacted directly. In principle these patches can be removed from the cleaning scripts after confirmation of amendments at the source level.

Aortic procedures

- *Affected records:* St. Thomas' Hospital 1st April 1999 – 31st March 2001 (approx.).
- *Fields affected:* 3.68-3.77 (inclusive), i.e. procedure and pathology fields for all five segments of the aorta.
- *Error:* all records indicate major aortic surgery (on all 5 segments). Patients are recorded as having an interposition tube graft for an aneurysm in in the root, ascending, arch and descending segments, and a dissection in the abdominal segment.
- *Resolution:* all records where the number of aortic segments operated on (3.67) was recorded as zero has the corresponding procedure and pathology fields wiped.
- *Notes:* does not resolve records where some segments were operated on but evidence all segments operated on. These records still correctly indicate a major aortic procedure, however identifying which segment the procedure was performed on and/or whether it was a thoracic procedure is not possible.

Valve procedures

- *Affected records:* St. Thomas' Hospital 1st April 1999 – 31st March 2008.
- *Fields affected:* 3.27-3.30 (inclusive), i.e. native valve pathology fields for all four-heart valves.
- *Error:* for any valve(s) having undergone a cardiac operation, the native valve pathology of the valves not operated on was recorded as '0. Native valve not present'.
- *Resolution:* a revised version of a valve procedure evidence flag (see Flags section) that omitted native valve pathology as an indicator was generated for each valve. For each valve, the native valve pathology was wiped for any revised evidence flag marked as 'False'.
- *Notes:* St. Thomas' were aware of this issue for 1st April 2008 – 31st March 2011 period after previous data validation exercises and subsequently resolved the issues. No attempt to revise previous and/or future records has been made.

Removed Records

Records were removed on four criteria:

1. If the dates were in conflict or before 01/01/1998.
2. If the patient was below the age of 18.
3. If the procedure (post-cleaning) was an isolated abdominal aorta procedure.
4. If the record was identified as a duplicate.

All records that are removed at any stage during the cleaning process are saved to an external CSV file for verification by database managers and clinicians at the appropriate hospital.

Date Conflicts



All records where the procedure date or discharge date is before 01/01/1998 are removed from the database due to the poor quality of the data prior to this time.

Only records that satisfy both the following two requirements are kept:

1. Admission Date \leq Procedure Date;
2. Procedure Date \leq Discharge Date.

If a comparison could not be made due to missingness, then the record was retained. Clerical error was generally unacceptable here due to the need to assess whether a record is a first cardiac operation in a specific admission window.

Adult Status

Any record where the patient age is less than 18 years at time of procedure (including patients aged 17 years and 11 months) are removed from the database since they are not adults by law.

Non-cardiac Procedures

Any record satisfying all of the following conditions were removed from the database:

- 3.11.1, 3.11.2 and 3.11.4 were recorded as '0. No'
- 3.11.3 recorded as '1. Yes'
- 3.69.1, 3.71.1, 3.73.1 and 3.75.1 recorded as missing
- 3.77.1 recorded as non-missing
- '3.67 Number of Aortic Segments operated on' recorded as 1

This is because these only provided information on isolated abdominal aortic surgery.

Duplicate Records

Where duplicate records exist, it is usually associated with subtle differences such as the first operator being listed differently.

We identified duplicated procedures by first extracting the following fields: '1.01 Hospital', '1.07 Gender', '2.36 Number of Previous Heart Operations', '3.11.1 CABG', '3.11.2 Valve', '3.11.3 Major aortic', '3.11.4 Cardiac procedures other', '3.01 Admission Date', '3.02 Procedure Date', '3.02 Procedure Time', '4.06 Discharge Date', 'Apollo', 'Age at operation'. In addition, an indicator was created: was procedure elective or non-elective. Any records identically matching on the above extracted fields are marked and the first record in the sequence kept.

ONS Data Merge

The ONS data was merged by deterministically matching on the common 'ParentUNID' fields. Not all patient records in the V4.1.2 database matched to a record in the ONS database; hence 'LS Date' and 'LS Status' were missing for a large number of records. This was primarily limited to Scottish and Irish patients. If multiple ONS records for a patient existed, the most recent record is used for merging.



Following the creation of flag variables (see next section), we checked whether there were conflicts between the database and the ONS data.

Records were **removed** if 'LS Date' preceded '3.02 Procedure Date' and 'LS Status' Date is 'Dead'. In this case the record was deleted. This is based on the principle that dead patients cannot undergo a first-time cardiac procedure.

Record linkage data were **altered** if 'LS Date' preceded '3.02 Procedure Date'. In this case 'LS Date' and 'LS Status' were set as missing. This is based on the assumption that the life status here is not a follow-up as intended.

Flags

New fields were created to provide evidence of a risk factor or patient status when multiple fields are related yet some have missing data or are in conflict with one another.

LV Ejection Fraction

This is an indicator of left ventricular ejection fraction category that resolves two fields: '2.27 Ejection Fraction' and '2.28 LV Ejection Fraction Category'

- If 2.28 was recorded, 'LVEFC flag' was set identically.
- If 2.28 was not recorded or unknown (i.e. missing or recorded as option '9. Not measured') but 2.27 was recorded, then 'LVEFC flag' was set appropriately.

Dead Flag

This is an indicator of in-hospital mortality status that resolves two fields: '4.04 Discharge Destination' and '4.05 Status at Discharge'

- If either 4.04 or 4.05 indicate the patient is dead and the other field does not contradict it (e.g. it is missing), then 'Dead flag' was set as 'True'.
- If either 4.04 or 4.05 indicate the patient is alive and other field does not contradict it (e.g. missing), then 'Dead flag' was set as 'False'.
- If there are contradictions between 4.04 and 4.05, 'Dead flag' was set as missing.

Dead Flag 2

This is an indicator of in-hospital mortality status that extends 'Dead flag' by using ONS census data to backfill missing and contradicting records.

For all 'Dead flag' defined as missing:

- If 'LS Date' matches '4.06 Discharge Date' and 'LS Status' is recorded as 'Dead', then 'Dead flag 2' is updated to 'True'.



- If 'LS Date' proceeds 4.06 and 'LS Status' is recorded as 'Alive', then 'Dead flag 2' is updated to 'False'.

Valve Procedures

This is an indicator of whether there is any evidence of a specific valve procedure occurring. Therefore we have four separate indicators: aortic valve flag, mitral valve flag, tricuspid valve flag and pulmonary valve flag.

For each valve in questions, evidence was equivalent to any of the following fields being non-empty

- Explant: 3.23, 3.24, 3.25 and 3.26.
- Pathology: 3.27, 3.28, 3.29 and 3.30.
- Replacement: 3.35, 3.36, 3.37 and 3.38.
- Procedure: 3.43, 3.44, 3.45 and 3.46.
- Implant type: 3.47, 3.48, 3.49, and 3.50.

Single Episode

This is an indicator of whether the record corresponds the (i) the only cardiac procedure for the admission spell, or (ii) the first cardiac procedure during the admission spell (i.e. where a patient had >1 cardiac procedure). The algorithm used to discern this flag was as follows.

- By default all 'Single episode flags' are set to 'True'.
- For patients with an Apollo number (England and Wales NHS records primarily):
 - If the Apollo number is only recorded once, then 'Single episode flag' remains unchanged;
 - If the Apollo number is recorded > 1 times, then records are ordered in ascending order according to Apollo number, then procedure date, procedure time, admission date, then discharge date.
 - If any record matches the previous record (for fixed Apollo number) for both admission and discharge date (or for just one if there is a missing date for the other), then 'Single episode flag' is recorded as 'False'.
 - If both the admission dates and the discharge dates cannot be evaluated as being equal (due to missing data), then:
 - if the procedure date is missing, then 'Single episode flag' is recorded as 'missing'; else
 - if the previous discharge date is not missing and the procedure date is less than or equal to this, 'Single episode flag' is recorded as 'False'; else
 - if the previous procedure date is not missing, then and the procedure date is:
 - within 2 weeks of this⁵, 'Single episode flag' is recorded as 'False'; or
 - greater than 1 year, 'Single episode flag' is recorded as 'True'; else
 - if the previous discharge date and the previous procedure date are missing, then 'Single episode flag' is recorded as 'missing'.

⁵ The choice of 14 days as a threshold is arbitrary and may not account for some patients who remain in hospital for long periods of time.



- For patients without an Apollo number (Scotland, Northern Ireland, Republic of Ireland and private hospital records mainly), but who have a ParentUNID value:
 - If the ParentUNID number is only recorded once, then 'Single episode flag' remains unchanged;
 - If the ParentUNID number is recorded > 1 times, then records are ordered in ascending order according to ParentUNID number, then procedure date, procedure time, admission date, then discharge date.
 - If any record matches the previous record (for fixed ParentUNID number) for both admission and discharge date (or for just one if there is a missing date for the other), then 'Single episode flag' is recorded as 'False'.
 - If both the admission dates and the discharge dates cannot be evaluated as being equal (due to missing data), then:
 - if the procedure date is missing, then 'Single episode flag' is recorded as 'missing'; else
 - if the previous discharge date is not missing and the procedure date is less than or equal to this, 'Single episode flag' is recorded as 'False'; else
 - if the previous procedure date is not missing, then and the procedure date is:
 - within 2 weeks of this⁶, 'Single episode flag' is recorded as 'False'; or
 - greater than 1 year, 'Single episode flag' is recorded as 'True'; else
 - if the previous discharge date and the previous procedure date are missing, then 'Single episode flag' is recorded as 'missing'.

Previous Cardiac Operation

This is an indicator of whether a patient has previously undergone cardiac surgery.

- 'Previous operation flag' was recorded as 'True' if any of the following were true:
 - '2.07 Previous Cardiac Interventions' contains options 1-5 (inclusive).
 - '2.36 Number of Previous Heart Operations' is recorded as > 0.
 - 'Single episode flag' is recorded as 'False' and is not missing.
 - Any of the 'Reason for Repeat valve replacement' fields (3.35, 3.36, 3.37 and 3.38) contained a non-missing response.
- 'Previous operation flag' was recorded as 'True' if both of the following were true:
 - '2.07 Previous Cardiac Interventions' was recorded as anything other than options 1-5 (which includes missing) and '2.36 Number of Previous Heart Operations' is recorded as 0 or missing.
 - All of the 'Reason for Repeat valve replacement' fields (3.35, 3.36, 3.37 and 3.38) were missing (i.e. not recorded due to not being applicable).

First Time Cardiac Procedure

This is an indicator of whether the record corresponds to the patient's first-ever cardiac operation. It was simply set as the opposite of 'Previous Cardiac Operation'.

CABG Flag⁷

⁶ The choice of 14 days as a threshold is arbitrary and may not account for some patients who remain in hospital for long periods of time.

⁷ This field resolves conflicts and missingness including the primary indicator. The principal only extends to CABG and valve, as per clinical judgement on typical data entry.



This is an indicator of whether there is any evidence to suggest a CABG procedure took place. It is recorded as 'True' if any of the following true:

- '3.11.1 CABG' is recorded as '1. Yes'. (This is the primary indicator used by researchers but is occasionally in conflict with the CABG specific data.)
- '3.14 Number of Grafts' > 0 and not missing.
- '3.15 Graft site' is not missing.
- '3.16 Graft conduit' is not missing.
- '3.17 Graft Anastomoses' is not missing.

Valve Flag

This is an indicator of whether there is any evidence a valve procedure took place. It is recorded as 'True' if any of the following true:

- '3.11.2 Valve' recorded as '1. Yes'. (This is the primary indicator used by researchers but is occasionally in conflict with the valve specific data.)
- '3.18 Number of Valves Repaired or Replaced' > 0 and not missing.
- Aortic valve flag recorded as 'True'.
- Mitral valve flag recorded as 'True'.
- Tricuspid valve flag recorded as 'True'.
- Pulmonary valve flag recorded as 'True'.

EuroSCORE

Three logistic EuroSCORE predictions are calculated here:

1. Original EuroSCORE
2. Modified EuroSCORE (SCTS internal model)
3. EuroSCORE II

In each model the covariates were generated for application in the model and two predictions generated: 1) where the prediction is missing if any covariate is missing; and 2) where each prediction is calculated under the assumption that any missing covariate is equivalent to the risk factor being missing or at baseline level. The postfix '.default' is appended to the second prediction fieldname.

The modified EuroSCORE design matrix (non-imputed) is appended to the database to facilitate external database users.

Shortcuts

As part of analysing the SCTS database certain routine fields are often required. A number of common variables have been generated and appended to the database after cleaning. These include:

- Country (England, Wales, Scotland, Northern Ireland and Republic of Ireland).
- Financial year.



- Geographical region (South West, South East, South Central, London, East England, West Midlands, East Midlands, North West, Yorkshire & Humber, North East, Scotland, Northern Ireland, Republic of Ireland, Wales).

Database Issues & Requests

In the process of cleaning the SCTS database we have identified multiple issues that require closer examination.

High Priority

- Hospital discharge status and ONS census records conflict. This is likely due to an error in the assignment of the unique patient identifier (ParentUNID).
- Records identified during cleaning as first-time valve procedures (based on rules described in the Flags section) were found to have conflicting inputs in the valve explant and pathology fields.

Medium Priority

- Resolve hospital stay temporal conflicts; i.e. admission should precede procedure which should procedure discharge.
- All issues listed under the Patches section.

Low Priority

- The field 'Age at operation' may be incorrect for borderline cases.
- Make available the records removed during the database cleaning for inspection by individual units.
- Cannot map peripheral vascular procedures from V3.8 '3.13 Other thoracic and vascular procedures' to '3.12 Other Cardiac Procedures' since original option coupled with aortic procedures.