Chapter 3: Data Validation Procedures

Introduction

It is the appropriate time to publish a description of the data validation procedures of the Renal Registry. Not only has the experience of data analysis from previous years reached a significant level, but it is also opportune in the year in which centres are identified in the report to demonstrate the care that the Registry has taken to ensure the quality of the data held. This chapter explains some of the extensive validation processes run by the Registry to make sure that the data are as accurate as possible.

The data validation process at the Renal Registry is divided into four separate parts:

1. Consistency checks when each of the quarterly files is loaded
2. Validation against previous data on the database after the quarterly file has been loaded
3. Validation on the database for consistency of data for the whole of the year
4. Statistical routines written in the SAS statistical language that run other consistency checks.

Processes 1 and 4 were in place at the end of the original 2 year pilot project but were often performed at a late stage. Delays in the Registry report resulted particularly from finding errors in the data at the fourth step. As a consequence, it was necessary to discuss problems with the affected renal units, eliminate any local database issues, re-run extractions, reload the Registry database, revalidate and then re-extract the data into SAS, a process that could add months to the reporting timescale. Over the past year, the Registry has increased the number of processes at stages 2 and 3 to be performed quarterly as data are received in order to reduce the number of potential errors picked up at a late stage.

Common data errors

The most common errors resulting in calls to renal units from the data management team are listed below:

1. Missing EDTA disease primary disease codes
2. Missing haemodialysis supervision code (i.e. home, hospital or satellite) against a treatment modality
3. Incorrect figures, e.g. blood pressure 130/4 mmHg
4. Dates that do not match, e.g. the date of starting dialysis being earlier than the date of birth
5. Checks on discrepancies relating to patient names and dates of birth between transplant and dialysis centres. These problems were particularly common when patients transferred for transplantation
6. Text in a numeric data field.

The Renal Registry modality timeline

The modality timeline contains information on the changes of modality. It is the most important item in the data sent to the Registry as the local extraction software uses this information to decide whether or not to send data to the Registry.

1. Local software detects renal replacement therapy (RRT) modality entries in this field to flag the patient as being on RRT and sends data to the Registry.
2. If a patient recovers function, the software sends no more data to the Registry for that patient until a new modality is detected (so ‘recovered function’ must not be used for transplanted patients).
3. If a patient ‘transfers out’, the software sends no more data to the Registry for that patient until the patient is transferred back in.

Accuracy is essential as analysis of these data by the Registry is critical in the following areas:

1. Take-on rates are calculated from the first modality in the timeline.
2. If the first modality is ‘transferred in’, the patient is excluded from the incident calculations.
3. Recovery of function excludes these patients from subsequent prevalence figures.
4. Prevalence rates are calculated by excluding any patients transferred out.
5. If patients are not recorded as having transferred out, no further results thus being obtained, they will be included in the analysis as having data missing.

The validation process

From the experience of the first 5 years, the Registry staff have recognised many common sources of error and have developed automated processes to check for these errors as early as possible in the data-loading process in order to avoid delays in data analysis after the data for the last quarter of the year have been loaded.

Consistency checks at load

Range tests for numeric items (e.g. biochemistry) and all the numeric data have simple range checks on them. These ranges have to be very wide to accommodate both physiological ranges (creatinine = 130 micromoles/L) and the pathological range (creatinine = 3500 micromoles/L). They also have to accept paediatric values.

The most frequent errors are found in the manual entry fields as a result of mis-keying the data. Many sites have not set up local range validation within their own database; instituting this feature would reduce the number of calls received from the data management team at the Registry. The most frequent errors are found in the blood pressure data, where for example an entry of 14 instead of 140 might be recorded. The second most common error is length of time on dialysis, hours being entered instead of minutes or, in other systems, minutes instead of hours. Similar mis-keying has been found in many other data fields, particularly where satellites units manually enter laboratory data.

We encourage sites, wherever possible, to add these range checks locally.

1. *Validity of coded entry fields.* Whenever possible, the Renal Registry uses Read codes or EDTA codes (for primary diagnosis and cause of death), these being validated against their acceptable ranges. Some renal units have modified the European EDTA codes, which this can cause validation errors.

   Again, a lack of fixed coding in the local renal unit database is the source of these errors. Some renal units allow manual entry of the EDTA code number rather than a specific diagnosis that is internally converted by the database into an EDTA code number. Sites that offered the ICD10 list of diagnoses for the primary diagnosis field have been successfully discouraged from this practice as this can result in errors – on one occasion, cerebrovascular accident was given as the primary cause of chronic renal failure.

2. *Avoidance of duplicate patients.* As patient data are received from both dialysis and transplant centres, it can be very easy, if the right systems are not in place, for patients to be duplicated on the database. This is not just a simple process of flagging up patients with the same surname, forename and date of birth. Many patients have their names spelt in a slightly different way on different databases, and it is impossible to impose consistency between two sites. In addition, dates of birth can vary (by days or months), and renal units have been unwilling to change these data, partly perhaps because the local automated laboratory links may refuse to load patient data if the date of birth or name is spelt differently from that held in the laboratory systems.

   Detecting these duplicated patients relies on a complex series of procedures:

1. *The NHS number.* This is a unique identifier for patients in patients in England & Wales, but many sites do not store this number in their renal system so it has not yet been found to be useful. The Registry submits files to the NHS number tracing service.

2. *UK Transplant number.* Patients on the transplant waiting list, or who have been transplanted, are allocated a unique number that should be stored on the local renal unit database (although it may be held only at the transplanting centre). This is recorded by the Registry as a non-duplicated indexed field. The Registry validates these numbers annually with UK Transplant, this process detecting mis-keyed data entry errors (e.g. 97074 instead of 90774). Renal units are informed of any mis-keying errors found.

3. *Soundex database index on names.*

4. *Check on the address field.*

   Once a new patient has been identified, the Registry allocates its own unique ID number for the patient, which is transmitted back to the renal unit and loaded electronically to avoid any mis-keying.

5. *Validation of the postcode.* The Renal Registry uses a commercial postcoding package (QAS systems), which checks the validity of the postcode against the address fields, the software automatically correcting any postcode errors. This is important as postcodes are used for NHS number tracing and also by the Registry for health authority mapping and deprivation scoring.

6. *Date fields.* All date fields must be in the format dd/mm/yyyy and are checked for validity.

7. There are other consistency checks, some of which are listed below:

    1) Date of first RRT < date of birth (as mis-keying has been found)

    2) Date of death < date of birth and < date of first RRT (as mis-keying of the date of death has occurred)

    3) Data are loaded for a patient already registered as dead. A transplant unit may have accidentally logged a patient as dead when he or she has returned to dialysis

    4) Mandatory fields (e.g. cause of primary renal disease) missing

    5) Inappropriate data for specific treatment modalities. A urea reduction ratio value or length of time on haemodialysis cannot be sent while the patient is on peritoneal dialysis

    6) There must be new patients starting RRT in every quarterly file received from the renal unit

    7) There must be some deaths in every quarterly file.

Consistency checks after loading

There are many errors that cannot be checked at the time the file is loaded. This is because these errors relate to the overall data integrity for the renal unit compared with the data from previous quarter.

1. Many long-standing transplant patients are not seen quarterly. Missing creatinine data over three consecutive quarters result in a telephone call to the unit in case the patient has died or transferred elsewhere without this being logged in the renal unit database.
2. Incorrect modality in a timeline. Some patients have had a modality of 'recovered renal function', rather than 'graft functioning', sent to the Registry after transplantation. This is invalid but can only be checked once the data have been loaded. Renal units are contacted to change this item.
3. The date of death must be later than last modality entry in the timeline. This item produced an error for 28 patients on one annual check, all caused by mis-keying the date of death.

Annual data checks

1. The number of patients starting RRT should be spread across all quarters, and the total number must be consistent with previous years.
2. Prevalent patient totals should rise on an annual basis.
3. Health authority take-on rates should show consistency.

4. In every centre, there should be timeline entries with a treatment start date for every quarter over which the Registry collected data (Table 3.1). There will always be some patients in a quarter who have changed modality; lack of modality changes indicates either a software problem or missing data.

Quarter no.	No. of new timeline entries
12	45
13	52
14	42
15	50
16	66
17	71
18	52
19	87
20	42
21	36
22	52
23	45
24	1

Table 3.1: Frequency of timeline entries per quarter – an example

5. Deaths should occur in every quarter, and the total number of deaths should be evenly spread. A lack of deaths registered often represents a software extraction fault rather than an error of logging by the renal unit. An excess of deaths in a given quarter can also be identified and of course investigated.

When errors such as those listed are found, the Registry contacts the renal unit for clarification and correction where appropriate.

SAS analysis

Many of these potential errors are rechecked through SAS, outliers being identified. Examples include the following:

1. *Have centres transferred patients in and out?* (The number of transfers in and out is usually roughly equal.) The results shown in Table 3.2 indicate that some dialysis centres do not transfer patients out prior to transplantation at another centre, and some centres do not transfer them back in. In such cases, the Registry adds the appropriate code to the timeline in the Registry database. This is important as, during the analysis, each patient is allocated to a centre for a specific time point. This is not only for calculating the total number of patients at a centre, but also for allocating missing biochemistry data to a centre.

Centre	Transfer in	Transfer out
RAJ01	50	61
RAQ01	181	195
RAZ	122	79
RCJAT	441	441
RCSLB	203	98
REE01	316	251
RF201	403	349
RFBAK	82	58
RFPFG	8	15

Centre	Transfer in	Transfer out
RGD03	2	
RH641	140	157
RH8	171	116
RHW01	29	48
RJ121	91	94
RK7CC	89	65
RK950	152	56
RKB01	190	207
RKHA4	134	210
RL403	169	149
RL7	411	396
RLGAY	146	147
RLNGH	274	234
RMF01	242	306
RNA03	200	232
RNX02	401	188
RQHC7	113	260
RQR13	1149	1118
RQS01		1
RRBBV	353	230

Table 3.2: Frequency of timeline transfer

2. *How many patients at each centre recover renal function during the first 360 days of starting RRT?* All hospitals should have a few patients who recover renal function during the course of their first year. If the number for a particular centre is too low, some patients are probably being misclassified as 'acute'. Conversely, centres with an excessive number of patients recovering within the first 90 days are probably including a large number with acute renal failure. Patients may be correctly classified as having chronic renal failure and recover function within 90 days.

3. *What are the first treatment modalities, and do any patients have two different treatment modalities on the same day of their first modality?*

4. *Have any patients been transplanted at a 'dialysis' centre?* Some dialysis centres do not transfer patients out for a transplant and furthermore may log the transplant as having taken place at their centre.

5. *Is there a variation in biochemistry results?* Software errors have been picked up by identifying centres if patients' results have not changed on a quarterly basis (e.g. haemoglobin level 10.6 g/dL) for two sequential quarters.

6. *Is there a difference between the pre-dialysis and post-dialysis blood pressure?* Such a difference is usually seen, so if the systolic and diastolic pressures are identical, one value is likely to be invalid.

7. *Has there been a large increase in missing data within a quarter for any data item*? This will be queried with the site. One site showed a large increase in missing urea reduction ratio data that had arisen from an undetected clerical error in storage of the post-dialysis sample data in the local database. Similarly, a reduction in returns for ethnicity was shown to be due to a change in the method of coding this item.

8. *Is there a complete absence of an expected biochemistry item (e.g. ferritin)?* This will be confirmed with the site. At two sites, the internal software links to these tables in the database were shown to be faulty.

9. The haemoglobin data for one site this year had begun arriving as an integer (e.g. 9 or 10 g/dL). This error will pass all the other validation checks and was shown up only on analysis of the distribution of these results (quartiles). This resulted in a software fix at the site and the necessity of re-exporting all the patient quarterly haemoglobin data for 2001. This was reloaded into the database and into SAS, and all the haemoglobin analyses were re-run (as the overall means will change) and graphs replaced in the text.

Comment

This is not an exhaustive description of the data validation routines, but the complexity of the task is evident. Much of the data quality depends on the monitoring of data entry at the originating centre. The local verification of data on a regular basis, through review meetings or weekly or monthly printed reports, is a necessary component of quality to assure the entries. The careful induction of new renal staff is another important issue. These checks may also be supplemented by automated routines for missing data and anomalous entries, a process with which the Registry is willing to help. A renal unit's progressive attention to the detail of data registration should allow a reduction in the validation effort required of Registry staff and a shorter interval between download and the production of the Report.

The key to further improvement in data quality is an open exchange of information and an increasing scrupulousness at renal unit level, which the UK Registry always attempts to match.