



Divisions of General Practice

Information Management Maturity Framework
(IMMF)

Toolkit – Technology solutions for
data validation and cleansing tool



Information Management Maturity Framework (IMMF)

Toolkit – Technology solutions for data validation and cleansing tool

Purpose

The purpose of the “Technology solutions for data validation and cleansing” tool is to assist Divisions to address the action task below.

Action Task	Capacity Gap	IMMF Element
Establish an inventory of current and desired specialist IM technology / tools available to facilitate the data validation process	Reactive to Defined	Compliance and Quality

This task should have been identified from the Information Management Maturity Framework (IMMF) gap analysis and toolkit specification.

This tool provides advice for Chief Executive Officers (CEOs) who wish to understand their options for using technology to assess the quality of data collected by the Division to support its own business activity and, where necessary, cleanse or remediate the data. It will assist the Division to meet its obligations to collect and report information collected from general practitioners (GPs).

This tool may be used in conjunction with the earlier tool, “Guidelines for data validation” which should be regarded as pre-requisite knowledge.

Explanatory notes

Data validation is important in maintaining efficient and effective operations within the Division itself and also, when field data are collected for aggregation across Divisions to assess the delivery and impact of major strategic health initiative programs on general practices and their patients.

Maintenance of high quality data collected from general practices depends on sound client relationships between a Division and its client GPs and their staff and, conversely, maintaining sound client relationships depends on many factors, including carefully planned and executed data collection strategies.

This tool acknowledges and relies on extensive earlier work commissioned by the Department of Health and Ageing in March 2007, that surveyed technology solutions for data validation in general practice. This and other references are attached at the end of this document.

Instructional design

This tool consists of three Parts:

Part 1 – Principles and factors driving data quality

Part 2 – Data validation – technology solutions

Part 3 – Cleansing data



Part 1 Principles and factors driving data quality

This part presents a brief discussion of data quality fundamentals and looks at the main factors that determine data quality.

Part 2 Data Validation – Technology Solutions

Part 2 focuses on a sample of technology-based approaches in use by some Divisions to assist them to improve their data validation processes.

Part 3 Cleansing data

Cleansing poor quality data is not an easily automated task. It is typically undertaken as a manual process, targeting one or more specifically identified problem areas. This part of the tool outlines sound practices in dealing with poor quality data.

Summary of outcomes and resources

Workstreams	Outcomes	Resources
<p>New processes or procedures to be adopted</p>	<p>Data validation processes and standards exist to ensure that the Division’s information assets are accurate, consistent, complete and current. These standards are applied to all the Division’s programs and services.</p>	<p>This tool is mentored for the implementation of new processes and procedures by senior staff from other Divisions.</p> <p>Knowledge and skills transfer for staff is expected to be implemented through group workshops.</p>
<p>Technology to be developed or acquired</p>	<p>Some tools are available to automate the data validation process in some of the Division’s programs and services.</p>	<p>Technology transfer is anticipated to require onsite visits for implementation up to 5 days.</p>



Part 1: Principles and Factors Driving Data Quality

1.1 What is data quality?

Data quality refers to a collective assessment of data based on an ideal where data have the characteristics of being:

- appropriate - the data set truly reflects that which it is expected to describe;
- complete - the degree to which the sample includes all the expected values;
- accurate - the degree to which the recorded data reflects the actuality in the field; and
- timely - there has been little or no significant change in field values since the time when the data were recorded.

Data quality is not an absolute concept. Data quality should be understood in terms of fitness for purpose. For example, a sample of five percent from a large and comparatively homogenous patient population may provide sufficient data points to support statistically valid inferences from a dataset. However, a client database containing contact details for 95% of the GPs in a Division might be considered to be deficient for the purpose of maintaining sound client relationships (where complete coverage might be expected). In turn, the term “fitness for purpose” may be proscribed or may be a matter of professional judgment.

This raises the matter of “tolerance”, which refers to the willingness of the manager of the data to accept incompleteness, inaccuracy, or data that may be out of date to some extent.

Deficiency in any one of the above data characteristics has the potential to devalue the data. However, unlike the dataset characteristics of completeness, accuracy and timeliness (which can be improved through additional data collection), a dataset found to be inappropriate is worthless and must be discarded.

An incomplete individual data set may be less problematic when these data are intended to be aggregated with many other data sets. However, if the data is to be used, for example, to address communications with a practice, a single missing data item can prevent a Division from providing services to a client.

Similarly, the question of whether a mature data set may be considered to be current depends on the expected rate of change of its elements. For example, whether there has been significant change in the environment from which the data were originally collected and a subjective assessment of the rigour of the initial collection process. It is common through the passage of time to see changes in data definitions in required data elements and refinement of the collection processes. While this may improve the utility of the data, it can also defeat longitudinal comparisons.

1.2 What factors influence data quality?

There are three key factors influencing data quality:

1. Clear definition of data sets and data elements within them.
2. Consistent collection processes that recognise “fitness for purpose”.
3. Careful validation (and, if necessary, cleansing) before data are entered into electronic systems, or aggregated into larger datasets, or incorporated into summary statistics.

When data are collected from different sources, each will have an attendant error rate. Aggregation of data from many sources leads to a summation of these errors. It stands to reason, then, that the most effective way of maintaining data quality in a database, is to carefully manage the data at the point of collection and entry and to segregate data to ensure aggregation from many sources and collating like with like.

Having well understood standard processes and procedures and having experienced people collecting the data are key pre-requisites. However, much of the data managed by Divisions have been drawn from GP practice systems where data may be less well-defined and data entry might not be so consistently rigorous. Moreover, much data collection requires transposition from paper records and this opens up the potential for an additional source of error. It is important to recognise that the entry of clinical data in a practice frequently has the single purpose of being an aide memoire for clinicians, and that maintaining consistency to facilitate external aggregation of data may be a serendipitous rather than a constantly-achieved outcome. It is



estimated that only between 10% and 20% of GPs who use computers to collect clinical data also do so regularly for producing collaborative data sets or reporting on specific patient groups.

Inconsistent quality control for much data at source suggests that data validation work in a Division should aim for maintaining the quality of aggregated data held in the Division, to a level that is “adequate for purpose” more so than to aim for total accuracy.

Validating the data at time of entry into general practice systems is undoubtedly the best and most efficient means of maintaining quality. In doing so, the Division will reduce the need for validation tools and the time and resources related to costly manual inspection of data, rectification at source and rework in a second collation.

Improving the quality of data entry in general practices suggests the importance and utility of a Division providing assistance and training of practice staff to promote this outcome. The value of providing training and support can also be seen as an investment that reduces the need for later remediation at higher cost.



Part 2: Data Validation – Technology Solutions

2.1 What is data validation using technology solutions?

Validating data using technology solutions involves the collection of data directly into spreadsheets or databases. Unique data (for example, a practice's contact information) may be manually collected and entered into electronic or paper-based systems by Division staff, by GPs or their staff. Alternatively, the data may be extracted from the systems in use within GP practices.

Extraction tools are provided by major GP practice systems vendors and some Divisions have invested considerable effort in extending reporting and extraction capability, for the benefit of both the Division and the practices they serve.

The following list is indicative of initiatives undertaken by Divisions. It is not exhaustive, but describes the nation's most frequently used tools and programs:

- Practice Health Atlas- particularly in SA Divisions. The Practice Health Atlas (PHA) is a decision support tool, designed by the Adelaide Western General Practice Network, for GPs, practice managers and other practice staff. The PHA is provided as a complimentary service to members and aims to inspire general practice teams to reflect on their activities and to develop business models for more effective health care services / outcomes. It is based on the synthesis of relevant, high quality and timely practice health data, as well as using such data to predict future health care needs and trends.
- National Primary Care Collaboratives (NPCC) / Australian Primary Care Collaboratives (APCC) Program provides a generic quality improvement model that can be applied to achieve incremental, rapid and locally relevant improvements across a broad range of clinical and practice business issues. Participating practices submit monthly data to an online data reporting system and are in turn provided analysis and feedback on the impact of their improvements over time. The Canning Tool, described below, has been used extensively for data extraction.
- Canning Tool is often used for NPCC / APCC and NPI data extraction and sometimes for use as a clinical audit tool. The Canning data extraction tool was developed by the Canning Division of General Practice to extract data from general practice management systems. The data is extracted into a set of tables which enables a Division to comply with the Department of Health and Ageing's National Performance Indicator (NPI) reporting requirements and also, the APCC reporting requirement. The Canning Tool is widely used in many Divisions across Australia.
- Pen Clinical Audit Tool - is a clinical information system that supports quality improvement in information management and enhances the business capability of general practice. It is a software tool that operates with the GP Clinical Desktop System to present the GP and other practice staff with meaningful clinical information. The Pen Tool presents aggregated patient information of the practice and encourages accuracy.
- Practice Nurse Initiatives - The Southern Division of South Australia changed its constitution to allow practice managers and practice nurses to become full members of the Division. Practice nurses are also recruited and paid for over a period to facilitate the relationship with the practice and to drive improved collection and management of clinical data in practice software.
- Enhanced Divisional Quality Use of Medicines (EDQUM) and National Prescribing Service (NPS) initiatives. A special version of Canning Tools was developed to assist the EDQUM program. This program re-invests in public health activities in the Division with savings produced in the cost of the Pharmaceutical Benefits Scheme as a result of the Division's efforts to improve the use of medications.
- Quality Use of Administrative and Clinical Software (QUACS). This is an initiative of the Inner Eastern Melbourne Division of General Practice to improve practice information management and data quality in particular, through better use of Practice Management systems. While focused on Medical Director, it gives a broader view of the features of practice management systems and how they can be used to improve data quality in the practice, to improve productivity.



- Whitehorse Division – Clinical audit program “Reflective Practice” is also based on the PHA and the Canning Tool. This involves working with the National Prescribing Service (NPS) clinical audits, which are provided free of charge to GPs and attract Royal Australian College of General Practitioners (RACGP) continuing professional development (clinical audit) points. Data are extracted using the Canning data extraction tool and can be displayed using the PHA to identify geographic patterns of illness and prescribing. NPS case studies are included in NPS News and mailed to all GPs every two months with the Australian Prescriber.
- North West Melbourne Division ran a pilot project to enable general practices to remotely access patient records in aged care facilities, in order to ensure that patient data are kept up to date. The aged care home was connected with the GP clinic by computer, which made it possible to establish which patients were resident in the home, to collect data, and then to synchronise the data between the clinic and the aged care home.
- Macarthur – Diabetes management. The Division is re-designing its current diabetes systems to include closer collaboration with the local health service and the development of its “Cardiab” database to better meet the combined needs of the Division and the Macarthur Diabetes Service. It involves the development of alternate data collection systems to improve the capture of diabetes data and simultaneously reduce the input requirements at the GP end. Better data quality will enable diabetes patients to be more easily identified, enabling them to be regularly recalled to have their condition more closely monitored, and to ensure they have greater access to continuity of care from diabetes specialists.

The list is not exhaustive, but reflects the commitment of Divisions in the change management process for improved collection and analysis of clinical data.

2.3 How is Automated Data Validation Useful?

Divisions are required to collect and summarise large volumes of information about the work of general practice. For summaries to be valid, the great majority of source data must be sound and data validation (and more particularly, cleansing of poor quality data) needs to be efficient and effective. Faster validation processes mean that more resources can be directed into data cleansing and into initiatives that promote better data management at source.

Automating parts of the data validation process assists by:

- reducing the time taken to scan data sets for gross errors;
- promoting a level of precision in identifying potentially problematic data that is difficult to achieve manually;
- allowing staff undertaking data validation to develop a feel for the data by producing statistics showing central tendency and variation (these statistics form one means of rapidly identifying outlying and suspect data); and
- dividing data validation work into a two step process – a fast scan followed by a closer examination of suspicious data.

2.4 What are the Features and Benefits of Automating Data Validation with Computer Technology?

Maintaining data quality is at the heart of the IMMF, for understanding and assisting general practice and for collating statistics and performance indicators that reflect the work of the Divisions, Department of Health and Ageing and other agency programs.

However, there is an inverse relationship between the need for effort in maintaining data quality (particularly at source) and the quality of data itself. Divisions with GP client bases with less experience in optimal use of computer-based Practice Management systems, have a difficult choice in deciding whether to concentrate data assurance resources on promoting good data collection practices up front, on detecting poor quality data and correcting the data itself, and addressing the root cause.



The main features and benefits of automating data validation with computer technology are:

- accommodating a large data volume and finding the needles amongst the hay;
- potentially achieving greater precision in detecting problematic data;
- exposing system problems;
- improving efficiency of the validation processes; and
- freeing data assurance resources for data cleansing or remediation work.

2.5 Choosing Which Data to Validate – and How Much Validation to Do

The choice amongst data validation methodologies also depends on the notion of “materiality” which is shorthand for a judgement of the anticipated impact of poor quality data. Materiality takes into consideration questions such as:

1. Operational business or policy consequences of the data being incorrect or of an otherwise poor quality.
2. Whether there are legal implications or financial penalties for failure of compliance with legislative or contractual obligations.
3. The frequency that the data are used. This might mean the number of transactions and their size or the number of individuals impacted.
4. The significance of decisions based on the data, especially financial decisions.

Data validation is often a multi-step process where a small sample of data is closely examined, and depending on whether the quality is within acceptable levels of tolerance (the examination of a pre-agreed sample size yields more or less errors than a benchmark), a decision may be made to extend the sample, or in fact test all the data for one or more characteristics. This recognises the common occurrence, that data that are problematic for example in timeliness, have a high probability of also being inaccurate, particularly if collected by different personnel or through different processes.

The results of the initial data validation exercise can then form the basis of a decision about appropriate steps needed to drive down the error rate. This is a cornerstone of Plan Do Study Act (PDSA) continuous improvement.

2.6 Analysing the Data – Identifying Potentially Problematic Data

Analysing data typically takes the form of one or more of the following:

- testing data against specific rules;
- checking peer group consistency – comparing similar practice statistics;
- checking historical trends for individual practices or practices within the Division;
- comparing practice or group statistics with published national benchmarks or historical data; and
- spotting data that ‘just doesn’t look right’ under visual examination.

In examining tabular data visually, missing data are obvious, as are data out of form (for example, text in a cell where a number is expected, or vice versa).

The analysis of the data can be automated to some extent and the mechanics of this are discussed at length in the “Guidelines for data validation tool”. In summary, basic validation mainly takes the form of automated scanning of columns (or rows) of data in tabular form.

Scanning is most quickly done by comparing data items with rule sets that ensure, for example, that a date record is collected in the correct format, and that a historical record cannot be assigned a date in the future.

One of the most comprehensive and accessible sources of benchmarking data is the Primary Health Care Research and Information Service (PHCRIS) website¹.

¹ See References – Primary Health Care Research and Information Service (PHC RIS)



2.7 Common Rule-Based Automatic Testing Includes

- **Format or picture checks**

Check that the data is in a specified format (template), e.g. dates have to be in the format DD/MM/YYYY.

- **Data type checks**

Check the data type of the input and give an error message if the input data does not match with the chosen data type, e.g. in an input box accepting numeric data, if the letter 'O' was typed instead of the number zero, an error message would appear.

- **Range check**

Checks that the data lay within a specified range of values, e.g. the month of a person's date of birth should lie between 1 and 12.

- **Limit check**

Unlike range checks, data is checked for one limit only, upper OR lower, e.g. data should not be greater than 2 (>2).

- **Presence check**

Checks that important data are actually present and have not been missed out, e.g. practices may be required to have their telephone numbers listed.

- **Spelling check**

Looks for spelling and grammar errors

- **Consistency Checks**

Checks fields to ensure data in these fields corresponds, e.g., If Title = "Mr.", then Gender = "M".

Automated data validation methodologies often rely on examination of a sample of the data, which can and sometimes does extend to all possible data points, meaning the whole of a discreet population.



2.8 Data Validation in Excel 2003 and 2007

Microsoft Excel 2003 allows the developer of a data validation spreadsheet to specify permissible values in cells when data are entered manually, but does not prevent or highlight invalid items. Excel 2007 has more extensive data validation capability. Users can easily specify the data format and permissible values that are expected in a range of cells. The function provides some flexibility about how non-compliant cell contents may be processed, including an option for simply placing a red circle around bad data to create a visual highlight.

2.9 Analysis Toolpak – an Add-in for Excel 2003 and 2007

In large tables with hundreds of values in a column, it is often beneficial to use the packaged statistical functions built into the spreadsheet. These generate several statistics that can be used to highlight unexpected data and also to estimate missing data items, if it is appropriate to do so.

Both Excel 2003 and Excel 2007 have instructions for using data analysis tools in the “Help” function. The most useful statistics are “descriptive” statistics – arithmetic mean, median (the middle score in a series), mode (most frequently observed score), minimum, maximum and range and standard deviation.

Unexpected values are likely to be those that are distant from any observable central tendency (mean, mode or median). Data that cluster closely around some value exhibit small variance and outlying values are comparatively easy to differentiate. In data sets where means and standard deviations can be calculated, 95% of values are expected to occur within two standard deviations above and below the mean and 98% of values are expected to lie within three standard deviations of the mean. Values outside of this range are not automatically understood to be incorrect, but they may well be worthy of checking to find some valid reason(s) why they have been observed.

When an observed data point is greatly different from a national statistic (based on a large number of values), there should be plausible reason. It is most likely to be due to either a particularly skewed demographic amongst patients attending a practice, or in aggregation, amongst numerous practices in a Division. Another frequent cause is an idiosyncratic use of a practice management package to store data in an unusual, if not inappropriate way for the purposes of aggregation.

Individual practices may develop approaches to enter data in package databases that provide additional functionality (for example some practices use allergy alert fields to store patient birthdays because they flash on screen and provide handy reminders). While this is useful for individual practices, it bodes ill for data aggregation by inflating the incidence of allergy. Similarly, some practice management software allows asthma data to be stored in four different locations which may not be used consistently and which confound accurate aggregation. This is one example of a major IM issue impacting practice data aggregation in many Divisions.

2.10 Analysing the Data – Chasing Down the Source of the Problem

- Revisiting the source.
- Protecting client’s confidentiality.
- Confirming that the received and sent data (or paper and transcribed computer records) are consistent.
- Checking the data at source – and identifying unconventional use of Practice Management packages.
- Dealing with missing data elements.

Two important questions:

1. Were the data captured consistently and categorised correctly amongst practices within the group and amongst individuals entering data with for example, a larger practice?
2. Were the data extracted and transmitted correctly?

General practices, as businesses, are naturally concerned about the collection of information that they could reasonably regard as being confidential and it is important to maintain this confidentiality (and be seen to do so) as a precursor to securing co-operation in keeping missing data to a minimum.



Divisional staff charged with the responsibility of pursuing missing data or dealing with a large number of clearly incorrect data from individual practices need to be cognisant of the stress and time needed to rectify the problem from the practices' perspective and attempt to approach practices with this in mind.



Part 3: Cleansing Data

Cleansing data is essentially a manual process, whether that data is textual material such as contact details or other information for a practice, used by the Division to provide its business services, or whether the data are to be used for internal reporting (Divisional key performance indicators) or external reporting (health programs and initiatives).

Cleansing data can take the form of a number of options, as follows:

1. Removing incorrect values (with or without replacement).
2. Finding the corresponding correct values and replacing the erroneous values.
3. Estimating a data element (noting that it is an estimate and recording how the estimate was made).
4. Segregating or re-classifying data that are judged to be too heterogenous to be meaningfully aggregated for use in like-with-like comparisons.

Amongst the automated tools reviewed, only the PHA specifically provides a data cleansing capability at the time this toolkit element was published.

The decision about whether it is preferable to work with a smaller data set or to insert estimates of missing data to keep sample sizes consistent depends on whether it is permissible in the context of the aggregation of that data.

There are several approaches to estimating missing numerical data. If there is a reasonable time series of data for preceding periods, it may be acceptable to substitute a calculated average value from preceding periods. Moving three-year averages are common choices. Alternatively, if data show no particular trend over time, a missing value may be filled with the mode (most commonly occurring) value or the mean value.

Option 2 above is the natural choice for non-numeric data.

In general, data cleansing is a function better minimised through prevention and good planning and it might be preferable, prior to a major data collection exercise, to thoroughly test the process from end to end with a small but representative sample of usually collaborative practices.



References

Bowers, David, Medical Statistics from Scratch, Second Edition, John Wiley and Sons, (2008)

BSR Solutions, Software Analysis Report, Appendix 1, (2007)

Canning Division of General Practice, Western Australia, Canning Data Extraction Tools

Available at: http://www.canningdivision.com.au/lnk_downloads.html

Last visited 7 August 2008

Pen Clinical Audit Tool

Available at: http://www.pencs.com.au/prod_detail.asp?cat_id=10&prod_id=20

Last visited 7 August 2008

Primary Health Care Research Information Service (PHC RIS)

Available at: <http://www.phcris.org.au> and <http://www.phcris.org.au/products/benchmarking.php>
(benchmarking data)

Last visited 7 August 2008

Practice Health Atlas

Available at: <http://www.awdgp.org.au/site/index.cfm?display=5462>

Last visited 7 August 2008

Spiegel, Murray, R., *Statistics*, Schaum's Outline Series, McGraw Hill, (2000)

Valintus Pty Ltd, Survey of State-based Organisations (2007)

End of Document