



Divisions of General Practice

Information Management Maturity Framework (IMMF)

Toolkit – Guidelines for data validation



Information Management Maturity Framework (IMMF)

Toolkit – Guidelines for data validation

Purpose

1. **The purpose of the “Guidelines for data validation” is to assist Divisions address the action tasks below:**

Action Tasks	Capacity Gap	IMMF Element
Implement a data validation process.	Unaware to Reactive	Compliance and Quality

2. **One or more of these tasks should have been identified from the Information Management Maturity Framework (IMMF) gap analysis and toolkit specification.**

This toolkit provides guidelines for confirming the accuracy and completeness of information records aligned to the Divisions programs and services for information management (IM). Divisions will be provided with an overview of the principles of data validation as well as guidelines to assist them in establishing a data validation process for any set of data being collected.

Knowledge of data validation principles and establishing data validation processes as it applies to Divisions is a pre-requisite for access to more advanced tools including the tool: “Technology Solutions for Data Validation and Cleansing Tools”.

Explanatory notes

Data validation is an integral component of Step 2, “Information collection and capture” which is addressed in the IMMF Toolkit - “Training pack for the information lifecycle (ILC)”. It provides a mechanism to develop a systematic approach for the collection of data by Divisions to ensure that data quality is high. Quality and validation relate to data having the characteristics of being accurate, consistent, complete and current. By adopting the guidelines defined in this tool, the Division’s information management doctrine will align with the IMMF.

A major benefit to Divisions in improving the quality of data is the reduction in time spent and cost of reviewing and fixing data errors prior to dissemination. It also means the end user of the data has a high level of confidence with the data quality.

By defining data validation processes, Chief Executive Officers (CEOs) will also have the ability to create common standards for data validation allowing them to share information and experiences with other Divisions.

It is the responsibility of all staff involved in the collection and capture of data to follow a process that improves the quality of the data. Support should be provided by the CEO to relevant staff to develop and implement data validation processes within the Division.

As a pre-requisite to developing data validation processes within the Division, senior staff must be familiar with the concept of an “information audit”. An information audit should be completed to identify areas that will benefit most from the investment in time and effort for data validation. This data would typically be classed as high value, for example data associated with funding programs.

The primary source of information was taken from “The CIHI Data Quality Framework”. A list of other references is documented at the end of this tool that provides useful information about data validation.

Instructional design

3. **This tool consists of two Parts:**
4. **Part 1 – Overview of data validation**
5. **Part 2 – Guidelines for implementing data validation processes**



6. Part 1 – Overview of data validation

This Part of the tool will assist Divisions in understanding the basics of data validation and how they fit into the information lifecycle and describe a generic data validation process.

The goal of data validation is to make sure information is fit for purpose, based on defined requirements for completeness, accuracy, consistency and currency. This will enable appropriate checks to be incorporated into the data collection and capture process.

The CEO should ensure staff involved in the collection of data should be familiar with the principles described in this section prior to developing the data validation process for the selected data sets.

Qualified State Based Organisation (SBO) staff may be available to assist in providing advice on the skills and processes other Divisions have used for data validation activities.

7. Part 2 – Guidelines for implementing data validation processes

This section details the steps that should be performed in a formal data validation process. Each data collection or set of information should have its own data validation process with specific tasks and rules. Senior program staff should apply these processes accordingly, while other Division staff involved in the data collection should be able to monitor the output of the data validation processes. In addition to the guidelines, the tool also provides a number of practical examples of data validation scenarios.

CEOs should use these guidelines, and the examples, to create a data validation process within the Division and to pilot that process with one of the Division’s programs or services deemed to be handling high value data or information.

Qualified staff from other Divisions or SBO staff may be available to provide mentoring in the development and implementation of a pilot data validation process.

Summary of outcomes and resources

Workstreams	Outcomes	Resources
<p>New processes or procedures to be adopted</p>	<p>New processes are developed for ensuring information is accurate, complete, consistent and current.</p> <p>Implement a data validation process as a pilot with one of the Divisions programs or services.</p>	<p>Mentored by senior staff from other Divisions.</p>
<p>Skills or knowledge acquisition requirements for staff</p>	<p>Staff familiar with principles of data validation and can apply them in their own work environment.</p>	<p>Facilitated by SBO staff.</p>



Part 1: Overview of Data Validation

What is Data Validation?

Once data has been collected, its quality must be verified before it is entered or captured in the Divisions records management system. Data validation forms part of the IMMF criteria – information quality. Data validation is the confirmation of the accuracy completeness, consistency and timeliness of data. A validation process can be implemented to perform checks against relevant data. Checks may be as simple as validating dates, numbers and codes used in a data set or more complex checks that incorporate conditional checking on multiple fields for consistency. Any corrupt data should be capable of being identified so that it can be corrected.

How is it useful for Divisions?

Divisions are responsible for collecting, verifying and disseminating a broad range of data including reporting to the Department of Health and Ageing on National Performance Indicators in various priority areas such as Chronic Disease Management, Prevention and Early Intervention and Access.

By identifying high value target areas of data, Divisions can maximise their investment in developing and implementing validation processes.

Benefits of implementing Data Validation Processes

- Reducing errors and inconsistencies within the data.
- Improve the quality of data to benefit from funding schemes.
- Ability to manage exception data.
- Provide accurate, consistent, complete and timely data to internal and external organisations.
- Decisions made based on validated data are more defensible.
- Ability to compare data across organisations.
- As the data becomes more accurate, less time and effort is spent on the validation process, ultimately saving time and money.

The Information Audit

Divisions should conduct an audit of their information resources before detailed planning for data validation. The audit should review all the existing data and information the Division currently manages. The aim of the audit is to identify information that has a high value associated with it to maximise the benefit of introducing data validation processes. This may include data that is collected based on funding agreements or data that helps strengthen relationships with GP practices.

The Process of Validating Data

The validation process may be used as a program for continuous improvement (a key principle of the IMMF). As errors are detected, fixes may be applied to the systems that provide the data to improve the data quality at the source.

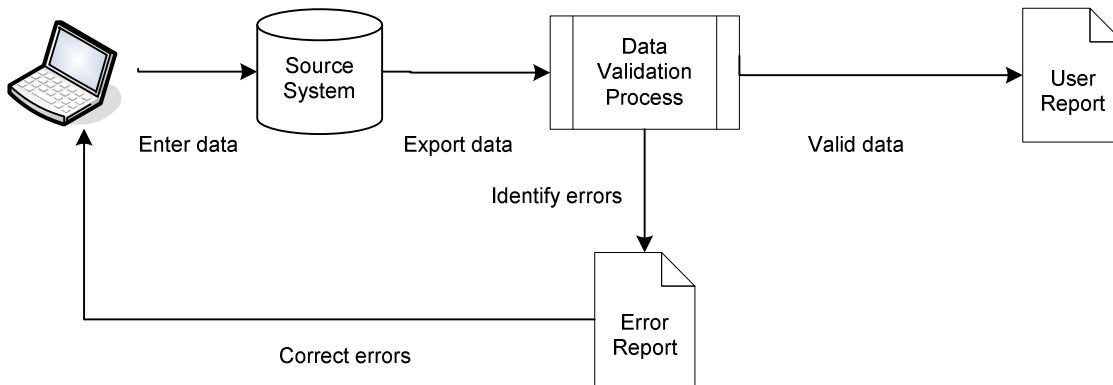


Figure 1

Figure 1 describes the processes used in a general data validation cycle. Data is entered into the source system. The source system will handle most of the validation as the data is entered. This is done by using business rules either through the user interface or data import routines (including manual data entry).

Further data validation may be applied to the source system or extracted data set to improve the data quality. An error report may be generated indicating the rows and/or columns that have data quality problems.

The error report may be used to correct data at the source system. It can also be used to identify new validation rules that may be implemented in the source system to improve the quality of the data as it is entered. It is preferable to correct errors at the source system in case any subsequent data extraction needs to take place and therefore minimising the need for potentially messy correction processes. It also ensures a high level of confidence with the data in the source system.

Once the data has been corrected it can be re-extracted and disseminated to relevant stakeholders.

Data Validation Rules

Data validation requires the creation of rules for different classes of data or for specific datasets. These rules have to cover the requirements for data quality in terms of:

- accuracy;
- completeness;
- consistency; and
- currency.

Hence a Division may create different rules for consistency for different data sets, e.g. the statistics for a general practice's diabetes register should be checked against national population statistics for diabetes; or date of birth fields for the use of the Red Book guidelines should be checked for the number of patients entered with 1st January birthdays.

Similar examples can be outlined for each of the factors of data quality. It is important that Division's address each factor with appropriate rules for each program or service within the organisation.

Sampling Methods

Sampling is generally used for statistical purposes but can also be used to determine how much data should be checked from a set of data to be comfortable of its quality. There are a number of sampling methods that can be used in the data validation process and the specific one used depends on how much time and resources are available and the "value" placed on the quality of the data.

The following list describes some sampling methods that may be used to check the quality of data within data sets:

- Comprehensive sample – this passes every record through the validation process. This method will provide the highest level of confidence of the validity of the data and will probably be used most of the time for the validation of high value data identified in the information audit.
- Simple random sample – records are taken at random from the data set and passed through the validation process.
- Systematic sample – this method involves looking at certain records at regular intervals, for example, every 10th record may be checked.



Automating the data validation process

The ultimate goal is to automate any validation processes that have been developed. Software is available to assist in performing this task and is discussed further in the IMMf tool - Technology Solutions for Data Validation and Cleansing Tools.

Steps to Implement a Pilot Data Validation Process

There are several steps that can be taken to establish a data validation process. These steps would follow an information audit. The idea is to establish the process so that it can be repeated on other sets of information.

1. Identify high value information that requires quality improvement.
2. Use the guidelines in Part 2 to define and develop components of the data validation process.
3. Test the validation process.
4. Refine the data validation components.

Key Resources

CEOs and senior staff should promote the introduction and development of data validation processes to all relevant staff within the Division. Senior staff should identify the areas of need and ensure the staff involved in these areas, understand the guidelines and how to implement data validation processes.



Part 2: Guidelines for implementing data validation processes

This section details the steps that should be performed in a formal data validation process. Each data collection or set of information should have its own data validation process with specific tasks and rules. The steps are not necessarily mandatory, depending on the data validation issues encountered it may be possible for the Divisions to eliminate some of the steps.

CEOs should use these guidelines and the examples provided, to create a data validation process within the Division and to pilot that process with one of the Division’s programs or services deemed to be handling high value data or information.

Step	Description
Information Audit	<p>The information audit is used to identify the value of the data that the Division collects. A plan can then be developed that identifies which data are the best candidates for implementing data validation processes against. These will be the higher value data.</p> <p>The following questions may be asked to determine high value data:</p> <ul style="list-style-type: none"> • Are there any financial incentives to collect the data? • Are the data used in a decision making process? • Can the data be used to build better relationships with clients?
Analysis	<p>It is important to understand as much as possible about the data that is being validated. In systems or database environments the main document that provides all of the information about the data is called a “Data Dictionary”. In less formal environments it could be a set of instructions on a form or a legend identifying responses to a question.</p> <p>The system that the data is being collected in has to be reviewed to identify any special characteristics or limitations that may be imposed on the data such as, field types or sizes. Often in databases a field is specified to a limited number of characters (e.g. 30) and longer strings of text are either truncated or lost.</p> <p>Finally, the use and method of dissemination of the data has to be considered.</p> <p>Each of these three factors has to be reviewed against the data validation requirements for accuracy, completeness, consistency and currency.</p>
Problem Description	<p>Document the problems that are occurring with the data in terms of the factors used in analysis. For example, data that includes a date field could be subject to problems that do not make sense from a business perspective. Here are some common problems:</p> <ul style="list-style-type: none"> • Accuracy - the date is outside the expected range. • Completeness - the field is mandatory and no value is present. • Consistency - the date cannot be after today (i.e. in the future). • Currency - the value reflects a previous data collection and has not been updated.



<p>Validation Checks</p>	<p>Validation definitions</p> <p>Validation checks/rules should be documented for each data item in question for accuracy, completeness, consistency and currency. An example of a template can be found at the end of this section.</p> <p>Data type validation</p> <p>This is one of the simplest checks to put in place. There are a number of different data types that may have been used and include string/character, numeric (integer and decimal), date/time, Boolean, etc.</p> <p>Checks can be developed to ensure that the data type for the column/field is valid. For example, checking that an integer field does not contain any alphabetic characters or that a date field contains a valid date.</p> <p>Mandatory data</p> <p>Sometimes a field or column is defined as being mandatory. This means that a value must be present. Checking for <i>null</i> or blank values can be implemented against the field.</p> <p>Range checking</p> <p>Range checking ensures that values contained in the field or column are within acceptable limits. A minimum and a maximum limit will be defined. Any values that fall outside the limits are deemed to be invalid. Range checking can be applied to any data type. For example, a date may need to fall between 1/1/2008 (minimum) and 30/6/2008 (maximum); a number may need to be between 1 (minimum) and 1000 (maximum).</p> <p>Code checking</p> <p>Code checking requires that values are checked against a set of 'codes' usually defined in the data definition. The code values may be stored in a lookup table. For example, the National Health Data Dictionary defines the data element <i>Sex</i> contains the values (1, 2, 3, 9). Any other values would be deemed invalid.</p> <p>Complex checking/business rules</p> <p>Data from different fields or columns may be conditionally related so that the value in one field determines what values can be valid in a different field. For example, a person's date of birth cannot be a date after their date of death.</p>
<p>Implement Error Check</p>	<p>The physical implementation of an error check can be performed using various tools. Examples of some tools that may be used include:</p> <ul style="list-style-type: none"> • Excel. • MS Access.
<p>Error Report</p>	<p>The error report is a product of the data validation process. It contains information about data errors that have been identified through the validation step. The report should indicate what error that has been found and where the error is located.</p>
<p>Data Correction</p>	<p>The information from the error report can be used to correct the errors, preferably at the source system. This ensures future extracts of the same data will be of higher quality and integrity. It also minimises the effort involved in rehandling the data.</p>
<p>Source System Correction</p>	<p>If errors are found to be consistently reported from the error report, business rules and edit checks should be identified and implemented in the source system. This will "fix" the error so it will no longer be reported from the validation process. The results of implementing this type of fix will be to improve the data quality and integrity at the source system reducing the time involved in data validation and cleansing.</p>



Using the Guidelines to Pilot a Data Validation Process

The following tables describe the process a Division may follow in order to implement a pilot validation process. They implement the guidelines for the data validation process and follow the Plan Do Study Act (PDSA) principles.

Three examples are provided to illustrating typical data validation scenarios faced by Divisions for:

- validating forms based data collection;
- validating clinical data; and
- validating KPI data.



Scenario 1 – Validating forms based data collection

Step	Description
Information audit	<p>The Division collects a range of information of varying quality. The collection methods range from telephone calls, paper forms and electronic submission (e.g. email, Excel spreadsheet and data files).</p> <p>The purpose of the data also varies for example, a paper survey may be sent to GP's to update their contact details or a collection of NPI data.</p> <p>The information audit should identify an area of high value to the Division and needs to start with some planning – for example: what is the audit meant to achieve? what are the acceptable standards for data quality? where might the data quality problems be?</p>
Analysis	<p>The instructions for collecting the data should be reviewed to determine if they answer the questions required for the required report and to ensure the values for each data item are appropriate. The data can then be data entered into an MS Access or similar database.</p>
Problem description	<p>Forms are sometimes sent back to the Division with incorrect and missing information. In many situations information problems are arising because the database cannot handle some of the responses. To address this problem, time needs to be spent to follow-up and correct the problematic information.</p> <p>Common errors that are occurring:</p> <ul style="list-style-type: none"> • date information is missing or not within the expected range; • responses to questions do not match the expected values (based on the associated instructions); and • characters have been entered into places where numbers were expected.
Validation checks	<p>Once the problem is identified, validation checks can now be defined and documented for the data items with data quality issues. The Division is now in a position to do something to improve the data quality such as:</p> <ul style="list-style-type: none"> • date: check that the date is within the prescribed rang; • codes: code checks will highlight those values that do not appear in the predefined set of values; and • data type checks: test for numeric values in the specified fields.
Implement error check	<p>The error checks can be implemented in MS Access or similar databases. The validation rules can be defined in a script that can be run to produce a report that contains records with errors.</p>
Error report	<p>The report can be generated from MS Access and similar databases to allow for easy review of the problems with the data. Studying the error report will provide the information to support action to improve the situation.</p>
Data correction	<p>The error report provides the details of the records that need attention. Staff can now follow-up and correct the data errors. The knowledge gained from this experience can be used to drive change in the collection process to try and reduce errors at the source.</p>
Source system correction	<p>Action will most likely need to be taken to improve the situation. For example, the data instructions may need to be reviewed to ensure clarity regarding the data items that have problems. Changes may also be made to the MS Access or similar databases to ensure all data is captured without modification.</p>



Scenario 2 – Clinical Data Audit

The following scenario describes the process a Division may follow in validate and correct clinical data at its source. It follows the principles embodied in the Australian Primary Care Collaboratives (APCC) of PDSA.

In this situation the Division has completed discussions with a member practice and that practice has agreed to work with the Division to take part in the APCC. A formal agreement has been made between the Division and the practice to access the practice data for the purpose of participating in the APCC.

Step	Description
Information audit	Division program staff work with the practice manager to extract from the practice information system an extract of patient records with a diagnosis of diabetes. The practice management system contains records on 1,563 patients. They plan to extract data to fit the purpose of the Australian Primary Care Collaborative (APCC) report and specify the diagnoses required should include Diabetes both Types 1 and 2. A data extraction software tool is used for the extraction.
Analysis	Jointly, the Division Chronic Disease Program officer and the practice manager analyse the extracted data to assess its accuracy, completeness and currency. The first thing they notice is that of the 1,563 patient records examined, there are only 33 patient records extracted that have a diagnosis of diabetes. The data extraction software tool they are using has the capacity to identify any variance in the percentage of patients extracted with a diabetes diagnosis from the national norm of 6-7% of the population. In this instance the extraction indicates that only around 2% of the patient population have diabetes. Given the practice deals with an older population in the area, the variance alerts the Division and the practice to a likely significant problem in the quality of the data. The practice manager discusses the variance with the GPs who identify several patients known to them that have not been identified as diabetic in the extract.
Problem description	Both the Division Diabetes Program Manager and the practice manager agree there is a significant problem of under-reported diabetes in the practice management system. A sample of individual patient records is examined and it is found that the diabetes diagnosis has frequently been entered in several different fields that the data extraction software is not counting and in some records the diagnosis has not been entered.
Data correction	<p>Jointly, the Division Diabetes Program Manager and the practice manager discuss the problem and plan what to do. They then take action to correct the data.</p> <p>A paper printout of the practice current patient list is made and the practice manager consults with all the GPs working in the practice who identify the current patients with a diagnosis of diabetes.</p> <p>The practice manager then accesses each of these patient records on the practice management system and enters the information into the correct field.</p> <p>The data extraction is then run again the following month and the results are compared to the previous extraction. This time the data shows that 125 patients have been reported with a diagnosis of diabetes. This is around 8% of the population, is close to the national norm and is the level expected for the population the practice serves.</p>
Procedure Changes	<p>The diabetes data quality issue is raised by the practice manager at the practice meeting and a short presentation is made on the correct fields in the practice management system to enter the diagnosis. This procedure is added to the practice procedure manual.</p> <p>The practice discusses the value of a recall system to ensure those patients identified with diabetes are systematically monitored and it is decided to implement the system.</p>



Scenario 3 – Validating KPI Data

The following scenario describes action taken by a Division to measure their success in engaging with member practices.

Level of contact with a practice during any given period is taken as a measure of member practice involvement with a Division.

Step	Description
Information audit	Practice liaison staff analyse the Division’s Customer Relationship Management (CRM) reports to identify member practices that would benefit by greater involvement with the Division’s programs and services. They run a report which shows the number of contacts over the past six months with each of the member practices in the Division.
Analysis	The report identifies four practices with several hundred contact records. This is over 10 times the average number of contacts with a member practice for the Division. Given the number of contacts with a practice has been designated as a key performance indicator, the practice liaison staff come to the conclusion that the data is not giving a true measure of Division relationships with those practices.
Problem description	Practice liaison staff then identify that the practices with the apparently excessive number of contacts are all practices which have a board member on the Division’s board and that board communications are being counted as practice contacts.
Data correction	Practice liaison staff meet with the Division CEO and it is agreed that a more appropriate way of gauging Division contacts would be through a combination of counts of program related visits to the practice and practice attendance at Division training activities. This information is recorded in different fields in the Division CRM. A new report is specified which counts those activities and this is discussed at a Division board meeting and is agreed to be a more accurate measure of practice involvement.
Procedure Changes	Key performance indicators are redefined to provide a more accurate measure of practice involvement with the Division.



References

IMMF Glossary

Data Quality Assessment – Communications of the ACM April 2002, Vol. 45.

NSW Health Data Quality Assessment March 2006

The CIHI Data Quality Framework June 2005 Revision

Primary Healthcare Research and Information Service
Available at: <http://www.phcris.org.au>

End of Document

