

# **The CIHI Data Quality Framework**

**June 2005 Revision**



Canadian Institute  
for Health Information

Institut canadien  
d'information sur la santé

All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or by any information storage and retrieval system now known or to be invented, without the prior permission in writing from the owner of the copyright, except by a reviewer who wishes to quote brief passages in connection with a review written for inclusion in a magazine, newspaper or broadcast.

Requests for permission should be addressed to:

Canadian Institute for Health Information  
495 Richmond Road  
Suite 600  
Ottawa, Ontario  
K2A 4H6

Telephone: (613) 241-7860  
Fax: (613) 241-8120  
[www.cihi.ca](http://www.cihi.ca)

© 2005 Canadian Institute for Health Information

Cette publication est disponible en français sous le titre : *Le cadre de la qualité des données de l'ICIS*  
*Révision de juin 2005*

# **The CIHI Data Quality Framework**

## **June 2005 Revision**

### **Table of Contents**

Acknowledgements .....	i
1. Introduction .....	1
2. The Data Quality Work Cycle .....	3
3. Assessment of Data Quality .....	5
3.1 The Assessment Tool .....	5
3.2 Table of Characteristics and Criteria by Dimension.....	7
4. Accuracy Dimension .....	9
4.1 Coverage .....	10
4.2 Capture and Collection.....	14
4.3 Unit Non-Response .....	16
4.4 Item (Partial) Non-Response .....	18
4.5 Measurement Error .....	20
4.6 Edit and Imputation .....	24
4.7 Processing and Estimation .....	26
5. Timeliness Dimension .....	29
5.1 Data Currency at the Time of Release .....	30
5.2 Documentation Currency .....	31
6. Comparability Dimension .....	33
6.1 Data Dictionary Standards .....	34
6.2 Standardization .....	35
6.3 Linkage .....	36
6.4 Equivalency .....	38
6.5 Historical Comparability.....	39
7. Usability Dimension.....	41
7.1 Accessibility .....	42
7.2 Documentation .....	43
7.3 Interpretability .....	44
8. Relevance Dimension .....	46
8.1 Adaptability .....	46
8.2 Value.....	47
9. Documentation.....	49
9.1 The Data Quality Assessment Report .....	49
9.2 Data Quality Documentation for Users .....	50
9.3 Methods Documentation .....	52
Appendix A—Glossary.....	53
Appendix B—Bibliography .....	63



# Acknowledgements

The Canadian Institute for Health Information (CIHI) would like to acknowledge and thank the many individuals and organizations—and in particular Statistics Canada—that have contributed to this document.



# 1. Introduction

The remarkable growth of CIHI has led to new and improved databases and registries, new users of CIHI information, new uses for CIHI data and a heightened visibility in the health information sector for CIHI. These increased responsibilities have prompted CIHI to introduce a data quality framework.

The data quality framework described herein embodies a systematic approach designed to apply to all databases and registries at CIHI. The framework organizes, assesses and incorporates activities related to data quality and is intended to be a living document that is subject to change on an ongoing basis.

## What Is Data Quality?

Although “data quality” is generally understood as a concept, the term is not well defined in current practice. The study of data quality as a topic is a relatively new field, and as such is complicated by differing definitions and approaches. Data is generally considered to be quality data (or to have sufficient data quality) when it would be appropriate to use for the purpose in question. The general consensus is that it is a difficult subject to measure—and no one measure is ideal for all situations.

## What Does the Framework Do?

The purpose of the framework is to aid in the systematic assessment, improvement and documentation of data quality for all databases and registries at CIHI. Given its objective, the framework is necessarily general in its design and deals primarily with issues that affect almost all databases and registries. The framework is intended to aid people who work on the databases and registries at CIHI in their ongoing efforts to identify areas needing improvement and in the resolution of any problems.

Data quality does not just happen. It can take a lot of work by different people to ensure that the data in a database or registry are as complete and as error-free as possible. Therefore, it is important to build the work and responsibilities related to data quality into the work plan of a database or registry. The first component of the data quality framework, the **data quality work cycle**, describes a work process that assigns roles and responsibilities and that identifies data quality priorities and the practice of continuous quality improvement. This is described in detail in Section 2, The Data Quality Work Cycle.

In order to ensure that the work cycle results in improvements to the data quality, it is important to periodically assess the quality of the data. The second component of the framework, the **assessment of data quality**, provides a way to assess the major characteristics of data quality. Although the concepts are all rigorously defined, they are intended to be general enough to apply to any database or registry. The assessment criteria are typically flexible and allow for adjustments due to the inherent subjectivity of many aspects of data quality. The assessment tool is described in Section 3, Assessment of Data Quality. The data quality concepts and assessment criteria are described in detail in Sections 4 to 8.

Informing users about the quality of data is arguably the most important part of a data quality program. It has been mentioned that data quality is subjective, and informing users about the quality of data allows those users to determine if the data are appropriate for their use. The third component of the data quality framework is the **documentation of data quality**. It is important to document the data quality for both internal and external users. Guidelines on documentation practices are in Section 9, Documentation.

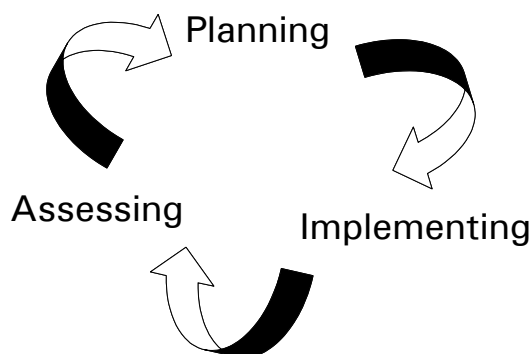


## 2. The Data Quality Work Cycle

The data quality framework is intended to provide a common objective approach to assessing data quality. It is therefore important that database and registry areas approach data quality in a common way. This includes not only using similar methods to assess the quality of the data and standard requirements for documentation, but also applying a consistent work process that identifies data quality priorities and produces continuous improvement in data quality.

Whether the end result of a work process is chocolate bars, automobiles or a database or registry, quality control experts recommend implementing a similar work process. Although the specifics of the processes may be different depending on the end product, the work process would involve an iterative cycle of ongoing assessment of the quality of the output and ongoing attempts to improve the quality of the output.

The following work cycle identifies a three-component approach that involves a set of planning, implementing and assessing activities, each of which are repeated as many times as needed throughout the cycle.



**Planning:** Includes the activities necessary to prepare and prioritize the processes required for a database or registry, as well as the design of any changes that are needed (such as deciding on the date of availability of the data and allocating appropriate resources).

**Implementing:** Includes developing the processes needed and applying them to the database or registry (such as collecting data, monitoring incoming records and releasing written reports).

**Assessing:** Involves evaluating the quality of the database or registry and determining if any changes to the processes are needed (such as completing a data quality assessment, writing data quality documentation and prioritizing required improvements). If any changes were to be developed, this development would take place during the planning stage. Thus, the cycle is iterative and continuous.

This three-stage cycle is designed to be flexible enough that it could apply to a database or registry at any point within its development life cycle. Due to the complicated nature of the databases at CIHI and the fact that databases typically operate on an annual cycle, the three components may overlap. Moreover, due to operational requirements, multiple ongoing initiatives, a long implementation period and unforeseen complications, the stages will often overlap. For example, it is more desirable to monitor the quality of incoming data on an ongoing basis than to wait until all data have been received to assess the quality. Ideally, analysis would be applied to the complete data set as well as to the data as they are received.

## **Roles and Responsibilities**

To ensure the success of the work cycle, it is important that everyone involved in the work process understand his or her roles and responsibilities. The Data Quality Coordinating Committee has established the roles and responsibilities below.

Senior management:

- Provide resources;
- Ensure that new initiatives incorporate the data quality framework from the feasibility stage forward; and
- Give overall direction and priorities.

Product areas:

- Analyze data to identify data quality issues and evaluate data quality;
- Address data quality issues when identified;
- Document data quality both for internal purposes and for users;
- Conduct special studies if appropriate; and
- Identify ways and means to improve the database or registry and its products.

The Data Quality section:

- Provide guidance on data quality methods and issues;
- Review and report on compliance with the data quality framework;
- Assist in special data quality studies in collaboration with the product areas;
- Conduct research and development activities relating to data quality methods and indicators;
- Provide educational and training services, as appropriate; and
- Update the data quality framework annually.

### 3. Assessment of Data Quality

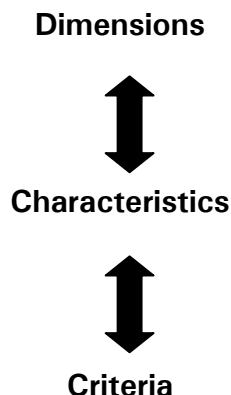
Assessment of data quality is accomplished through the use of the **assessment tool**, the core component of the framework. The tool identifies aspects of concern with relation to data quality for the database. CIHI's definition of data quality is **fitness for use**, which depends on the use being discussed and the standards of the user. The assessment tool will not attempt to measure fitness for use, but will attempt to identify aspects that have a negative effect on data quality so that these aspects can be accounted for and improved upon. Database strengths may also be identified to highlight processes that work well. Product areas will ultimately have the responsibility of determining whether their data are fit for use.

To create an operational definition of data quality, five dimensions of data quality have been defined to divide fitness for use into distinct components. The dimensions used at CIHI are accuracy, timeliness, comparability, usability and relevance and these will be expanded upon in the chapters to follow. The objective of separating data quality into components is to allow the identification of individual aspects that may create problems with regards to fitness for use. Although many organizations have split data quality into dimensions, there is no general consensus on what is the best way to clearly define data quality. Different organizations may use different dimensions.

The assessment tool's main purpose is to assess and document the limitations and strengths of a database. Typically, potential limitations will be identified by criteria rated as *not met*. Once these issues have been identified, they can be used to formulate recommendations for improvements. The assessment tool also captures the strong points of a database, allowing corporate best practices to be identified and shared between product areas.

#### 3.1 The Assessment Tool

The assessment tool is comprised of dimensions, characteristics and criteria. Each dimension is divided into related characteristics, and each characteristic is further made up of several criteria. These criteria are specific statements that, once assessed, help to identify the presence of potential data quality issues.



Dimensions are the distinct components that encompass the broader definition of data quality. They are:

### **Accuracy**

The accuracy dimension refers to how well information in or derived from the database or registry reflects the reality it was designed to measure.

### **Timeliness**

Timeliness refers primarily to how current or up to date the data are at the time of release, by measuring the gap between the end of the reference period to which the data pertain and the date on which the data become available to users.

### **Comparability**

The comparability dimension refers to the extent to which databases are consistent over time and use standard conventions (such as data elements or reporting periods), making them similar to other databases.

### **Usability**

Usability reflects the ease with which a database or registry's data may be understood and accessed.

### **Relevance**

Relevance reflects the degree to which a database or registry meets the current and potential future needs of users.

### **Ratings**

Ratings are used in the assessment tool as a guide to highlight strengths and identify weaknesses and limitations of a database or registry. Ratings are only one component for product areas to consider in determining fitness for use. The subjective nature of data quality and the differing nature of databases and registries mean that no rating system will be able to identify all problems with data quality. Scores which do not identify data quality issues do not necessarily mean a database or registry's data are problem free. Similarly, scores that indicate some data quality issues do not necessarily mean that a database or registry's data should not be used. It is the responsibility of the product area to identify in the presence or absence of data quality issues the strengths and weaknesses of the database or registry.

In most cases, each **criterion** is given a rating of either *met*, *not met*, *unknown* or *not applicable*. In select cases, criteria are rated according to other predetermined categories: *minimal or none*, *moderate*, *significant* or *unknown*. Each criterion has a statement or data table that can be used to determine what rating should be assigned.

These scores are for internal purposes only and should appear only in the evaluation report. They should not be included in any external documentation.

All assignments are made together with staff responsible for each database or registry. Determination of the level of quality must be done after confidentiality, privacy and security considerations have been addressed (if necessary).

### **Action Plan**

Once the assessment is completed, the next recommended step is to summarize the key findings from the assessment and develop an action plan that identifies strategies to remedy any deficiencies.

### 3.2 Table of Characteristics and Criteria by Dimension

Accuracy	
<b>Coverage</b>	1 The population of reference is explicitly stated in all releases
	2 Known sources of under- or over-coverage have been documented
	3 The frame has been validated by comparison with external and independent sources
	4 The rate of under- or over-coverage falls into one of the predefined categories
<b>Capture and Collection</b>	5 Practices exist that minimize response burden
	6 Practices exist that encourage cooperation
	7 Practices exist that give support to data suppliers
	8 Standard data submission forms and procedures exist
	9 Data capture quality control measures exist
<b>Unit Non-Response</b>	10 The magnitude of unit non-response is mentioned in the data quality documentation
	11 The number of records received is monitored to detect for unusual values
	12 The magnitude of unit non-response falls into one of the predetermined categories
<b>Item (Partial) Non-Response</b>	13 Item non-response is identified
	14 The magnitude of item non-response falls into one of the predetermined categories
<b>Measurement Error</b>	15 The level of measurement error falls into one of the predetermined categories
	16 The level of bias is not significant
	17 The degree of problems with consistency falls into one of the predetermined categories
<b>Edit and Imputation</b>	18 Validity checks are done for each data element
	19 Edit rules and imputation are logical and consistent
	20 Edit reports for users are easy to use and understand
	21 Imputation is automatically derived from edits
<b>Processing and Estimation</b>	22 Documentation for all data processes is maintained
	23 Documentation for all systems, programs or applications is maintained
	24 The processing system has been tested after the last revision
	25 Raw data are saved in a secure location
	26 The sampling bias and variance of the estimates are at acceptable levels

<b>Timeliness</b>	
<b>Data Currency at the Time of Release</b>	27 The difference between the actual date of release and the end of the reference period is reasonably brief 28 The official date of release was announced in advance of the release 29 The official date of release was met 30 Database or registry methods are regularly reviewed for efficiency
<b>Documentation Currency</b>	31 The recommended data quality documentation was available at the time of data or report release 32 Major database or registry reports were released on schedule
<b>Comparability</b>	
<b>Data Dictionary Standards</b>	33 Data elements are evaluated in comparison to the CIHI Data Dictionary 34 Data elements conform to the CIHI Data Dictionary
<b>Standardization</b>	35 Data are captured at the finest level of detail as is practical 36 For any derived data element, the original data element is also maintained on the main database
<b>Linkage</b>	37 Standard Geographical Classifications (SGC) can be used 38 Data are collected using a consistent time frame 39 Codes are used to uniquely identify institutions 40 Codes are used to uniquely identify persons
<b>Equivalency</b>	41 The impact of problems related to crosswalks or conversions falls into one of the predetermined categories 42 Methodology and crosswalks or conversions are documented
<b>Historical Comparability</b>	43 Trend analysis is used to examine changes in core data elements over time 44 The extent of problems in comparing data over time falls into one of the predetermined categories 45 Accessible documentation on historical changes to the database exists
<b>Usability</b>	
<b>Accessibility</b>	46 An official subset of microdata is defined, created, made available and frozen per release for users where appropriate 47 Standard tables and analyses are produced per release 48 Products are defined, catalogued and/or publicized
<b>Documentation</b>	49 Data quality documentation exists per annual subset release 50 Database or registry methods documentation exists for internal purposes per annual subset release 51 A caveat accompanies any official preliminary release
<b>Interpretability</b>	52 A mechanism is in place whereby key users can provide feedback to, and receive notice from, the product area 53 Revision guidelines are available and applied per annual subset release
<b>Relevance</b>	
<b>Adaptability</b>	54 Mechanisms are in place to keep clients and stakeholders informed of developments in the field 55 The database or registry can adapt to change
<b>Value</b>	56 The mandate of the data holding fills a health information gap 57 The level of usage of the data holding is monitored 58 User satisfaction is periodically solicited

## 4. Accuracy Dimension

*If you want data quality, prepare for bad data.*

Accuracy is what most people think of when they think of data quality. Accuracy refers to how well information in or derived from the database or registry reflects the reality it was designed to measure. When people ask if an estimate is good, most of the time they are asking if the estimate is accurate.

The accuracy of a database depends on many factors, some of which are difficult to measure. When considering accuracy, it is important to keep the following in mind:

- Are all the appropriate data present?

Three characteristics address this issue.

- Coverage—do you know who should be submitting data?
- Unit non-response—have all the records been submitted?
- Item non-response—are the submitted records complete?

- How good are the data?

Two characteristics address this issue.

- Capture and collection—what measures exist to minimize error?
- Measurement error—how well were the data reported to CIHI?

- What is done with the data?

Two characteristics address this issue.

- Edit and imputation—are the checks and modifications to the data logical and consistent?
- Estimation and processing—are the processes used to generate values documented and tested?

In this way, we have divided accuracy into seven characteristics. These characteristics are divided into 26 criteria, which are rated individually.

Dimension	Characteristics	Criteria
Accuracy	Coverage	1 to 4
	Capture and collection	5 to 9
	Unit non-response	10 to 12
	Item (partial) non-response	13 to 14
	Measurement error	15 to 17
	Edit and imputation	18 to 21
	Processing and estimation	22 to 26

## 4.1 Coverage

Criteria
1 <i>The population of reference is explicitly stated in all releases</i>
2 <i>Known sources of under- or over-coverage have been documented</i>
3 <i>The frame has been validated by comparison with external and independent sources</i>
4 <i>The rate of under- or over-coverage falls into one of the predefined categories</i>

Under- or over-coverage occurs when there is a difference between the **population of reference** and the **frame**.

The **population of interest** is the population for which information is wanted.

For example, the population of interest may be:

*All hospitals in Canada with at least one acute care bed.*

However, information for the complete population of interest is often not available. Instead, we must reference it using an available population: the population of reference. The **population of reference** is the population for which statements are made.

For example, the population of reference for a database may be:

*All publicly funded, non-prison hospitals with at least one acute care bed in all provinces and territories, except Saskatchewan and New Brunswick, that were open for business on January 1 of the reference year.*

The population of reference for a database is often complex, but should be coherent and consistent and, ideally, as close to the population of interest as possible.

The **frame** for a database is a list of provinces, institutions, doctors, etc. (units) that provides access to units in the population of reference that will be part of data collection. Although frames are primarily associated with sample surveys, administrative databases also need accurate frames. The frame should be used to determine from whom the data should be collected and what proportion of the data was actually received. Frames often have additional information about the units (for example, geographic location, contact information, number of beds) that can be used to stratify the population of reference.

**Under-coverage** occurs when part of the population of reference is not included on the frame that is used. For example, under-coverage would occur if some hospitals in British Columbia were not included when they should have been. Or if a health region in Nova Scotia was split into two health regions and only one was included on the frame.

**Over-coverage** occurs when units that are not part of the population of reference (that is, that are out of scope) are included on the frame or when duplicate records appear on the database.



Records that should not be included in the database or registry are known as **out-of-scope** records. For example, in the population of reference mentioned above, over-coverage would occur if any of the following three cases were included in the data: a hospital in Saskatchewan, an Ontario hospital with no acute care beds or a Quebec hospital that opened on March 1. Over-coverage would also occur if the same data were submitted by both the hospital and the province, resulting in duplicate records, or if a hospital submitted scheduled day surgery data to an emergency care database, resulting in the inclusion of out-of-scope records. These examples would result in over-coverage unless measures were taken to correct the data.

Because the frame is critical, it is important to ensure that **frame maintenance** occurs on a regular basis. Frame maintenance consists of adding any new units to the frame and removing any units that are no longer in the population of reference. All information pertaining to the frame should be kept up-to-date. **Frame maintenance procedures** are any practices that are used to update the frame.

It is important to realize that coverage errors, with the exception of duplicate records, do not necessarily relate to the submission of data. If a hospital on the frame does not submit data, this is an example of **non-response** and not under-coverage. For databases that use the frame only as a list of data suppliers, coverage errors often have to be detected by external verification (for example, list of suppliers from the provinces), but the errors that are detected are easy to correct (that is, the data suppliers are either added to or removed from the frame).

The degree of under- and over-coverage can be difficult to measure. Under-coverage in particular often has to be estimated.

For example, if we were collecting data on patients, the **rate of under-coverage** (expressed as a percentage) would be

$$\frac{\text{Units not on the frame but in the population of reference}}{\text{Units in the population of reference}} * 100\%$$

And the **rate of over-coverage** (expressed as a percentage) would be

$$\frac{\text{Units on the frame but not in the population of reference}}{\text{Units in the population of reference}} * 100\%$$

The units in the population of reference have to be adjusted for over- and under-coverage, as illustrated in the example below.

#### **Example: Calculating Coverage Error**

Number on the frame	=	1,100
Number missed	=	25
Number erroneously included	=	5

$$\begin{aligned}\text{Adjusted population} &= (\# \text{ on the frame}) + (\# \text{ missed}) - (\# \text{ erroneously included}) \\ &= 1,100 + 25 - 5 = 1,120\end{aligned}$$

$$\text{Under-coverage rate} = \frac{\# \text{ missed}}{\text{adjusted population}} * 100\% = \frac{25}{1,120} * 100\% = 2.2\%$$

$$\text{Over-coverage rate} = \frac{\# \text{ erroneously included}}{\text{adjusted population}} * 100\% = \frac{5}{1,120} * 100\% = 0.4\%$$

### **Criterion 1** *The population of reference is explicitly stated in all releases*

It is very important for the users of the data to know who or what is being examined; it is therefore important that the population of reference be explicitly mentioned in all releases. A release is any report, data release or output from the database or registry. Minor issues or clarifications may be mentioned in footnotes.

The population of reference for a release may be different from the population of reference for the database if the release involves a subset of the data (for example, Ontario data only). While it may be necessary to mention only the population of reference for that release, the population of reference for the database should be mentioned as well. The difference between the population of reference and the population of interest should also be noted.

This criterion is met only if the population of reference is stated in **all** releases during the last year.

### **Criterion 2** *Known sources of under- or over-coverage have been documented*

The known sources of under- or over-coverage should be documented on a regular basis. If the under- or over-coverage can be corrected, the data should be corrected before results are published and the source of error documented for internal use. Over-coverage can be corrected by removing the out-of-scope or duplicate records, while under-coverage is difficult to correct. If the data cannot be corrected, the sources of under- or over-coverage should be mentioned in all data quality documentation.

This criterion is met if the known sources of coverage error are documented internally or externally as required.

### **Criterion 3** *The frame has been validated by comparison with external and independent sources*

In order to detect errors on the frame, it may be necessary to compare the frame to sources external to and independent of the database or registry. Although frame maintenance should be done on a regular basis, it is necessary to compare the contents of the frame to an external source that is independent of the database's documentation (for example, a list of hospitals obtained from the provinces) to ensure the frame is up-to-date. In conducting any comparisons with external data, the credibility of the external source should be noted and, where possible, multiple sources used.

Data sources are generally considered independent if they are derived from different sources that are not related to the database in question. In many cases, finding an external source of information that is completely independent is not possible. For example, many of the data sources are derived from the same sources, such as provincial ministries of health. If an independent data source cannot be found, the frame should, at a minimum, be verified by the respective ministries of health. The ministries are a reliable external source of health information. If errors are found on the frame, it may be necessary to examine the frame maintenance procedures used.

It is important to note that if an external source cannot be found with the level of detail required, a comparison at an aggregate level can be done to detect errors on the frame. For example, if our population of reference was facilities with MRI machines in New Brunswick, and we could get only the number of facilities with MRI machines from an independent source, rather than a list of the facilities, we could compare the number of facilities on our frame to the independent number available from the external source. If the numbers match, that does not guarantee that there are no mistakes (as we may have the wrong facilities on our frame); if the numbers do not match, we know there is a problem with one of the sources of information.

This criterion is met if the frame has been compared to external sources within the last year to determine the presence of errors on the frame.

**Criterion 4**     *The rate of under- or over-coverage falls into one the predefined categories*

This criterion is designed to put some *qualitative* measure on the effect of under- or over-coverage. It is more subjective than the other criteria for this characteristic; however, it is possibly the most important. Consider the following guidelines:

Possible Rating	Rate of Under- or Over-coverage (%)
Minimal or none	Less than 1%
Moderate	1% to 5%
Significant	Greater than 5%
Unknown	Could not be determined

When deciding what rating to give a database, consider both under- and over-coverage separately and use whatever rating is worse. For example, a database with minimal over-coverage and significant under-coverage should be rated as *significant* for this criterion.

**These are only suggested ratings, as the effect of under- or over-coverage depends significantly on the amount and distribution of missing data.** For example, missing one hospital in Ontario may not affect provincial estimates significantly; however, missing a hospital in Prince Edward Island could drastically affect its provincial estimate. Also, it is important to realize that when dealing with patient data, one very large hospital can have the same effect as many small hospitals on under- or over-coverage.

It is very important to note that over-coverage does not compensate for under-coverage. For example, a database with 5% over-coverage and 5% under-coverage does not have good coverage. Out-of-scope or duplicate records do not compensate for missing records that should be present.

## 4.2 Capture and Collection

Criteria	
5	<i>Practices exist that minimize response burden</i>
6	<i>Practices exist that encourage cooperation</i>
7	<i>Practices exist that give support to data suppliers</i>
8	<i>Standard data submission forms and procedures exist</i>
9	<i>Data capture quality control measures exist</i>

The capture and collection characteristic refers to the practices that are used when dealing with the data suppliers and during data entry. The **data supplier** or **data provider** is the person or organization that provides the data to the database or registry. The data suppliers usually do the **data capture**, which is the actual entering of data into a usable format. **Data collection** is the gathering of the supplied data from different data providers into a common database or registry.

For a clinical example, the supplier may be defined as the hospital: data capture is the abstraction of hospital charts by the hospital, and collection is the receipt of the electronic abstracts by CIHI.

The relationship with data providers is of the utmost importance, as a good relationship not only increases the likelihood of response and the timeliness of response, but also the quality of the data. With regards to data suppliers, the question that should be asked is “What can we do to make the data supplier’s job as easy as possible, while still getting the information that we need?”

### Criterion 5 *Practices exist that minimize response burden*

This criterion assesses whether measures are used to ensure that the data supplier has to do as little unnecessary work as possible. This is otherwise known as minimizing the **response burden**. There are many ways this can be done: electronic capture and submission, auto-filling variables, reasonable submission schedules, exclusion of unnecessary data elements, etc. In cases where the data supplier or province collects data for its own purpose, the database area may have little control over what data elements are captured.

This criterion is met if practices are in place to minimize response burden.

**Criterion 6** *Practices exist that encourage cooperation*

Practices that encourage cooperation are important whether or not the submission of data is done on a voluntary basis, but especially when the data suppliers have the option of not responding. Practices can be as simple as stressing the importance of participation and the assurance of confidentiality, providing rewards for cooperation (for example, free publications, free training and specialized reports) or acknowledging the receipt of data in a letter or email and thanking the supplier for the information.

This criterion is met if there are any practices in place that encourage cooperation.

**Criterion 7** *Practices exist that give support to data suppliers*

Providing support to data suppliers is essential to ensure that data are submitted promptly and correctly. Technical and coding support should be made available to data suppliers. Support can include hotlines or an email address for questions, supporting documentation, guidelines for coding and education sessions.

This criterion is met if there are at least two methods by which support is given to the data suppliers.

**Criterion 8** *Standard data submission forms and procedures exist*

Standard data submission forms and procedures make the data collection process easier for both the data supplier and internal personnel at CIHI. It may also result in improved timeliness and reduced errors. It is not necessary to have only one data submission form, as separate language forms or separate forms for provinces may be necessary; however, the forms should be in a consistent format, with questions worded as consistently as possible. Standard submission procedures ensure that data collection is done as consistently as possible across suppliers. A database in which some institutions submit annual data on paper and others submit monthly data electronically does not have standard procedures.

This criterion is met if the data submission forms and procedures used in data collection are standardized.

**Criterion 9** *Data capture quality control measures exist*

**Data capture quality control** measures are carried out when the data capturers enter data. These measures ensure that the data are recorded properly. They can include data capture edit checks, visual verification of the data, dual capture or other procedures.

Data capture **edit checks** can greatly increase the quality of data sent to CIHI, as they allow verification or corrections while the original data are present. Data capture edit checks generally consist of validity checks by field (for example, checking to see if the gender is reported as M, F, Other or Unknown, or if the patient identification number is the appropriate length). Edit checks must also be applied when the data are loaded onto the database or registry (refer to section 4.6, Edit and Imputation).

**Visual verification** consists of having a second person examine the original data and the captured data for any differences, while **dual capture** is having two people independently record the data to check for differences. These procedures are often resource intensive, so they are frequently done on a sample basis.

This criterion is met if any quality control measures are used at the data capture stage.

## 4.3 Unit Non-Response

### Criteria

- 10 *The magnitude of unit non-response is mentioned in the data quality documentation*
- 11 *The number of records received is monitored to detect for unusual values*
- 12 *The magnitude of unit non-response falls into one of the predetermined categories*

**Unit non-response** occurs when entire records are missing from the database or registry. Unit non-response is often confused with under-coverage, as both occur when complete records are missing from the database. If a hospital on the frame for a database does not submit data, it is a case of unit non-response, whereas under-coverage would occur if the hospital was not on the frame.

The nature of administrative databases often results in a data collection process through data providers (such as hospitals) that allows non-response to occur at different levels. If hospitals are expected to submit data for their day-surgery patients, some hospitals may not submit any data, while other hospitals may submit records for only a portion of the day surgeries that take place there. These are both examples of non-response. It is important to get some measure of the amount of data missing from a database.

In CIHI, a **unit response rate** (the complement of the unit non-response rate) is usually computed rather than a unit non-response rate. The response rate can vary significantly if the calculation is based on the number of institutions or the number of records. The institution response rate (expressed as a percentage) is calculated as follows:

$$\frac{\text{\# of institutions that submitted data}}{\text{\# of institutions on the frame}} * 100\%$$

The response rate for the units of analysis at a particular hospital (expressed as a percentage) is:

$$\frac{\text{\# of records submitted by the institution}}{\text{\# of records that should have been submitted by the institution}} * 100\%$$

Depending on the data holding, it may be appropriate to calculate a response rate on sub-groups, such as size, type, region or province, in addition to calculating an overall response rate. Missing data from a large institution can result in the loss of many more records than missing data from a small institution. The approximate response rate for the records can be calculated by comparing the overall number of records received to the number of records expected. This calculation can become quite complex. Consultation with Data Quality personnel is recommended if such a calculation is done.

While it may be easy to detect institution non-response, it is not always possible to tell if the institutions have provided all the required records. It is, however, relatively easy to detect large changes in the number of records that an institution submits. Significant changes can give an indication of a low response rate or the inclusion of incorrect records and should be examined. The number of records received from each institution should be monitored on an ongoing basis so that unusual numbers can be examined.

**Criterion 10** *The magnitude of unit non-response is reported in the data quality documentation*

It is important that the magnitude of unit non-response be reported in the data quality documentation so that users can assess the completeness of the data. If response rates vary significantly by province or region, the response rate should be reported at these levels.

This criterion is met if the magnitude of unit non-response is reported in the data quality documentation at a level of detail relevant for most analysis.

**Criterion 11** *The number of records received is monitored to detect for unusual values*

In order to detect non-response below the frame level (for example, from the institution), it is important that the number of records be tracked over time to detect unusual values. For example, if an institution submits approximately 1,000 records monthly, the institution should be examined if the number of records received suddenly jumps to 1,500 or drops to 500. The change in the number of records submitted does not necessarily indicate problems, as there are many possible reasons for the change (for example, late submission of records, expansion of the institution or temporary closure).

This criterion is met if the numbers of records received below the frame level are monitored over time for unusual values.

**Criterion 12** *The magnitude of unit non-response falls into one of the predetermined categories*

As mentioned earlier, non-response may occur at several levels. This criterion considers the frame level (often institution). Basically, this means the data received are compared to the units on the frame. As indicated previously, a unit non-response rate which is the complement of the unit non-response rate is usually computed in CIHI rather than the unit non-response rate.

The unit response rate (expressed as a percentage) is:

$$\frac{\text{\# of units that submitted data}}{\text{\# of units on the frame}} * 100\%$$

The following table has suggested ratings dependent on the response rate at the frame level. If there is substantial non-response below the frame level, or the frame units that did not report are either very small or very large, the ratings can be adjusted to more accurately reflect the severity of the problems created by the missing data.

Suggested Rating	Unit Response Rate (%)
Minimal or none	Greater than 98% to 100%
Moderate	90% to 98%
Significant	Less than 90%

## 4.4 Item (Partial) Non-Response

### Criteria

*13 Item non-response is identified*

*14 The magnitude of item non-response falls into one of the predetermined categories*

**Item non-response** (or **partial non-response**, as it is sometimes known), occurs when a record that is received has some blank data elements that should not be blank. Item non-response differs from unit non-response, in that unit non-response deals with units or records that are missing, while item non-response deals with data elements that are missing when they should not be blank.

Item non-response rates are easy to calculate, if it is possible to identify when the data elements (that is, the variables) are missing. If we asked hospitals the number of psychiatric beds they have, a blank answer may mean that the institution has no psychiatric beds or that it simply did not respond. Data elements with blank values are normally identified (that is, flagged) during the editing process by the creation of an additional data element that identifies whether the blank value is acceptable or not.

In CIHI, an **item response rate** (the complement of the item non-response rate) is usually computed, rather than an item non-response rate.

The item response rate for a data element (expressed as a percentage) is:

$$\frac{\text{\# of data elements for which data was reported}}{\text{\# of reporting records that should have reported the data element}} * 100\%$$

For example, if 100 clinics were asked two questions: How many doctors work at your clinic? and How many patients were seen last month? Ninety of the clinics submitted data (resulting in a 90% **unit response rate**), with all 90 providing data on the number of doctors, but only 75 answering the question about the number of patients.



In this example, the item response rate for number of doctors is  $90/90 = 100\%$  (all clinics that submitted data submitted this data element), while the item response rate for the number of patients would be  $75/90 = 83.3\%$ .

It is important to realize that the item response rate is calculated in comparison to the number of records that *were* reported and not the number that *should have been* reported for the data element. In this example, although we have an 83.3% item response rate for the number of patients, we have patient data from only 75% (75/100) of the clinics. The item response rate does not give a full picture of the completeness of the data, so it is important to consider unit non-response as well. A database with a 100% item-response rate for all data elements may be missing a lot of data if the unit response rate is low.

Item response rate should be calculated for all data elements that are used in analysis of the data.

### **Criterion 13** *Item non-response is identified*

In order to determine the extent of item non-response in a database, it is important to be able to distinguish between blank values and non-response. Item non-response cannot be simply identified by a blank value. A flag should be used to identify that the value is missing when it should not be blank. Some data elements are required only conditionally (for example, if a patient is in for childbirth, it may be important to know if she previously had a child by caesarean section, but it may not be important otherwise). If these data elements are missing when required, they should be flagged as missing.

Data elements are normally flagged during the editing process by the creation of an additional value to the existing data element (for example, -1), or the creation of a new data element altogether. This allows one to easily identify whether the original data element is missing.

This criterion is considered met if item non-response can be identified for all **core data elements** and is flagged when the data element is required only conditionally. A core data element is one routinely used in analysis.

### **Criterion 14** *The magnitude of item non-response falls in one of the predetermined categories*

This criterion focuses only on the volume of data elements with item non-response. The effect of item non-response depends on many factors, including the importance of the missing data elements and whether there is any pattern to the missing values. Ideally, the missing values should be “missing at random” to avoid bias in the results. Data are considered missing at random if the values are missing on a random basis and not related to any other data elements.

The level of non-response should be assessed for each core data element. A core data element is one routinely used in analysis. As indicated previously, an item non-response rate which is the complement of the item non-response rate is usually computed in CIHI rather than the item non-response rate. When rating the level of item non-response, the

core data element with the lowest response rate should be considered. This may not provide an overall item non-response rate but, for the purposes of the assessment, will identify problem areas by establishing the lowest level of response. As these are suggested ratings, if the missing data elements are especially important or it is determined that they are not missing at random, the rating may be changed to more accurately reflect the severity of the problems created by the missing data.

The suggested ratings are:

Suggested Rating	Item Response Rate (%)
Minimal or none	Greater than 98% to 100%
Moderate	95% to 98%
Significant	Less than 95%

## 4.5 Measurement Error

Criteria
<i>15 The level of measurement error falls into one of the predetermined categories</i>
<i>16 The level of bias is not significant</i>
<i>17 The degree of problems with consistency falls into one of the predetermined categories</i>

The measurement, bias and consistency criteria combine to give a measure of how well the data were reported. An indication of the quality of the records submitted to CIHI can be given by the degree of coding error and the reliability of the data elements being coded.

Coding errors occur when the value reported to CIHI is different from what the value should be. Since errors can occur for many different reasons, it can be difficult to group the errors to allow an easy assessment of the differing causes. The framework divides the assessment of errors into three overlapping components:

1. **Measurement error** assesses to what degree the values reported match the values that should have been reported. This is similar to a validity check in epidemiological terminology.
2. **Bias** assesses to what degree the difference between the reported values and the values that should have been reported occurs in a systematic way.
3. **Consistency** assesses the amount of variation that would occur if repeated measurements were done.

**Measurement error** provides an indication of the number of times that a data element is coded incorrectly, while **bias** provides an indication of whether or not the errors that are present occurred in a systematic way. Although errors in a database are never beneficial, random errors provide significantly problems than errors that result from bias. While random errors may have little or no effect on the overall estimates of a database, errors that result from bias can dramatically affect the overall estimates.

For example, in looking at two scales: one consistently gives weights that are 2 kilograms too light; the other gives weights that vary a lot for repeated measurements. If someone used the first scale, they would consistently report values 2 kilograms lighter than the true weight. However, if they used the second scale, their measurements may vary if repeated, but on average, they would have a closer estimate to the true weight (if the scale did not tend to underestimate or overestimate weights).

The *existence* of bias can be very hard to prove without special studies, although it can be relatively easy to detect *possible* biases. For example: imagine that we examined the rates of Attention Deficit Hyperactivity Disorder (ADHD) by province and found upon further examination that Manitoba had an unusually high rate of ADHD in relation to the other provinces. This, by itself, is not necessarily an indication of bias, because ADHD may be more prevalent in Manitoba, or the doctors and teachers in Manitoba may be better able to detect ADHD. However, if we knew that Manitoba was the only province that distributes financial aid to the parents of ADHD children in order to allow for extra tutoring, this would be a situation in which bias could easily occur and the higher rates of ADHD would be a good indication that there may be a bias. However, it is important to note that this is not proof of a bias. Bias can result from under-reporting as well as over-reporting. In the ADHD example, the actual bias may be the under-reporting of ADHD from other provinces relative to Manitoba. Manitoba may be reporting accurately because there is a financial incentive to do so, or there may be a combination of under-reporting by other provinces and over-reporting by Manitoba.

When considering bias, it is important to consider **correlated bias**, which is a bias that is correlated with another data element (such as a province). The ADHD example above is an example of a correlated bias, because the bias is correlated with the province. Although correlated bias can be more complicated than uncorrelated bias, it is often easier to detect because the values can be compared across data elements (by province in the ADHD example or by scale in the weight example) and differences detected.

**Consistency** measures the variation of the responses over repeated measurements and, in some cases, is referred to as reliability. Subjective variables (such as level of impairment on a scale of 1 to 5, or diagnosis type) are data elements that may not have a correct answer. Consistency not only applies to subjective variables, but can also be a factor for data elements where there is an element of measurement error (for example, measuring height). The consistency characteristic provides insight into how much variation in the coding might be due to the differing opinions of coders. There are several statistical techniques that can measure the consistency of coding, such as percentage agreement or the kappa statistic. These estimates and tests of agreement among multiple coders are appropriate when responses are on a nominal or ordinal scale. Consult Data Quality staff for information on which technique is most appropriate for the data.

While the exact level of measurement error and bias and the consistency of data elements is often not known unless re-abstraction studies or other special studies are completed, the personnel working with the database often have some idea of the quality of the data elements in the database and can provide an initial assessment of these three factors.

**Criterion 15** *The level of measurement error falls into one of the predetermined categories*

The amount of error in the data elements of a database is most often assessed through re-abstraction or other special (and usually retrospective) studies, but it can also be assessed when the database is being developed. The level of error is often expressed as an error rate or a discrepancy rate and is measured by the percentage of cases for each data element that was coded incorrectly. Since this assumes a correct answer is possible, error rates do not typically apply to subjective variables.

If a re-abstraction type study has been done, the results for the core data elements should be examined and compared to the table below to get a suggested rating. The rating for the core data element with the highest error rate should be used, but if many data elements have substantial error rates, or the data elements are either especially important or non-important, the rating can be adjusted accordingly.

Suggested Rating	Error Rate for Non-Subjective Variables (%)
Minimal or none	0% to less than 5%
Moderate	5% to 10%
Significant	Greater than 10%

If the level of error is not estimated through a data quality study, that does not necessarily mean the criterion is automatically rated as *unknown*. In many cases, people working on the database may be aware of problems or at least have some idea of the amount of error in the data before special studies are conducted. This awareness must be used with supporting information to assign a level, rather than assigning a level based on a precise numeric error rate. If the level of error has not been assessed with a study and the database personnel are unable to evaluate it qualitatively, the criterion should be rated *unknown*.

**Criterion 16** *The level of bias is not significant*

The measurement error criterion assesses the amount of error in the non-subjective variables in the database, while the bias criterion is designed to assess whether the differences in the reported values are systematic. Bias, however, can apply to both non-subjective and subjective variables. For example, the response to a dentist's question of how many times a week a patient flosses his or her teeth (a non-subjective variable) is often biased, because people are more likely to exaggerate the number of times they floss to make the dentist happy. Similarly, if someone is trying to get compensation or sympathy for an injury, they will often exaggerate the amount of pain they are in, which is a subjective variable.

This criterion is difficult to evaluate, due to the complexity of proving whether a bias has occurred. The assessment, therefore, is based on whether there is, or is perceived to be, a substantial bias in the data. Biases (or possible biases) in the data can be detected by comparison of estimates to external sources, internal comparisons to detect correlated bias (values by province, hospital, etc.) and verification of records through re-abstraction.

If there is, or is believed to be, a bias (or correlated bias) in the data that is significant enough to affect the estimates to a noticeable degree, the level of bias should be rated as *not met*. If there is no evidence of bias and no reason to believe there is a bias, the level of bias should be rated as *met*. Otherwise, the criterion should be rated as *unknown*.

This criterion is met if there is no evidence of, and no reason to believe there is, a bias significant enough to affect the estimates to a noticeable degree.

**Criterion 17** *The degree of problems with consistency falls into one of the predetermined categories*

Consistency is a concern for all databases that have data elements that may depend on the opinion or interpretation of the coders. Consider both the consistency of the measurements from the individual coders, and the consistency of measurements between coders. For example, a classification specialist may interpret a chart the same way each time he or she looks at it, but different classification specialists may have differing interpretations. Similar to bias and error, consistency is most often assessed through re-abstraction studies, but it can also be assessed when the database is being developed. Random spot checks to measure consistency can be done throughout data collection. As discussed previously under this characteristic, there are several statistical techniques to measure the consistency and accuracy of coding. See Data Quality staff for information on the measurement technique that is most appropriate for the data in question.

If a special data quality study (a re-abstraction study, for example) has been done, the results for the subjective variables should be examined and compared to the table below to get a suggested rating. The rating for data elements with the lowest level of consistency should be used, but if many data elements have problems with consistency or the data elements are either especially important or non-important, the rating can be adjusted accordingly.

Suggested Rating	Discrepancy Rate (%)	Kappa Statistic
Minimal or none	0% to less than 5%	0.81 to 1.00
Moderate	5% to 10%	0.50 to 0.8
Significant	Greater than 10%	Less than 0.50

If the level of consistency is not estimated through a data quality study, that does not necessarily mean the criterion is automatically rated as *unknown*. In many cases, people working on the database may be aware of problems or at least have some idea of the reliability of the data before special studies are conducted. This awareness must be used with supporting information to assign a level, rather than assigning a level based on a precise numeric error rate. If the level of consistency has not been assessed with a study and the database personnel do not feel able to comment on it, the criterion should be rated as *unknown*.

## 4.6 Edit and Imputation

### Criteria

- 18 *Validity checks are done for each data element*
- 19 *Edit rules and imputation are logical and consistent*
- 20 *Edit reports for users are easy to use and understand*
- 21 *Imputation is automatically derived from edits*

Different organizations may use different distinctions between editing and imputation. For example, editing may refer to both the identification and modification of problematic data. In CIHI, **editing** is the process of identifying incorrect or missing data that should not be blank and **imputation** is the process of determining and assigning replacement values for incorrect or missing data that should not be blank identified at the editing stage. It is important to ensure that all data are valid and any changes made to the data do not adversely affect the quality of the data.

Editing of the data can be a complex task. Edits are typically divided into two types: hard and soft edits. **Hard edits** are used to identify when data are definitely incorrect, whereas **soft edits** are used to identify when data are possibly incorrect. For example, hard edits can be used to ensure that discharge date is later than admission date, because if the discharge date is before the admission date there is obviously an error in one of the fields. If most stays in a hospital are of a short duration, a soft edit may check to see if the discharge date is more than a year after the admission date. A long stay like this does not necessarily mean an error was made, just that the value is unusual enough to warrant attention. Data that are flagged by soft edits should be manually verified, if time allows.

Setting up proper edits is an investment in data quality. A good editing program can identify a lot of errors in the data that might not be detected otherwise. When setting up edits, it is important to consider whether it would be appropriate to use a hard or soft edit and what the parameters of the edit are. For example, how long does a stay in hospital have to be to warrant examination of the date fields? The above example of a year-long duration was chosen at random, but the value could be adjusted to try to maximize the improvement in data quality while limiting the amount of manual verification. If the duration parameter of the edit is too short, a lot of valid data will be validated manually, which can take up a lot of resources. If the duration parameter is too long, invalid data may not be flagged. It is recommended that you consult with your data quality personnel if you have any questions or are planning any complicated edits.

### Criterion 18 *Validity checks are done for each data element*

Validity checks ensure that the proper response format is used and the response is appropriate. Validity checks can consist of comparing the response to a list of acceptable responses (for example, a list of diagnosis codes) or simply ensuring that the response is in a proper format (for example, a four-digit code). A validity check on a date variable can be either to ensure that the response is in an acceptable date format (for example, yyyyymmdd), or to ensure that the response is an acceptable date.

It is important to realize that a valid response does not necessarily mean that the response is correct; it just means that the response *could* have happened. If a surgery is reported to have occurred on February 27, this is a valid date; but it may or may not be the correct date. However, if the surgery was reported to have occurred on February 31, the reported date is known to be incorrect.

Invalid data in a database will quickly raise questions about the quality of the data, and as invalid data are nearly impossible to justify (how would you explain a patient weighing -20 kg?), it is very important that invalid data be identified. Depending on the nature of the database, invalid data may be excluded, sent back to the supplier for correction, flagged for imputation, or simply flagged as invalid and dealt with separately.

This criterion is met if all collected data elements are checked for validity and any invalid data are flagged as invalid.

### **Criterion 19** *Edit rules and imputation are logical and consistent*

Edit rules are logical if they make sense with regards to the data that are collected. Automated imputation should not be used to modify data that may be correct, but can be used to modify data that are obviously incorrect.

Consistency edits are edits that are performed, in combination, across data elements. For example, consistency edits can be used to flag data such as a 70-year-old woman having a baby, a man having a caesarean section, a 4-year old with an occupation or a hospital with 50 nurses listing a total salary budget of less than \$100,000.

This criterion is met if the edit rules and imputation are determined to be logical, and if obvious consistency checks are in place.

### **Criterion 20** *Edit reports for users are easy to use and understand*

Edit reports should identify the records that passed or failed the edits and why they failed the edit rules. Any imputation should also be identified.

In order to have efficient editing and imputation processes, it is important that edit reports be easy to understand. In cases where the data are sent back to be modified by the data suppliers, it is especially important that the reason the records failed the edits be clearly reported to the data suppliers. If the reason is not given, it can be very time-consuming for the data suppliers to manually examine these records to determine what modifications are needed.

The number of times data are imputed or the number of times a particular edit rule is used has implications about the quality of the data. If an edit rule is used more than should be expected, it may mean that the edit rule is too restrictive or that the incoming data is of unusually poor quality. In either case, the data and the edit rule should be examined. This criterion is met if the edit reports are easy to understand and in a usable format.

### **Criterion 21** *Imputation is automatically derived from edits*

Imputation is the process of determining and assigning replacement values for incorrect or missing data that should not be blank identified at the editing stage. Making any changes to the data submitted by the data providers is considered imputation.

Although imputation is an accepted statistical practice when properly done and accounted for, manual imputation is often subjective, difficult to trace, easily questioned (why one value and not another?) and not supported by any sound statistical theory. Proper imputation, on the other hand, is less subjective, easy to trace, difficult to question when done appropriately and supported by statistical theory and practice.

Imputation can be very complicated and can have drastic effects on the data quality if done improperly; therefore, it is suggested that any imputation schemes be developed with the assistance of a statistician or methodologist.

This criterion is met if the imputation process is automated.

## **4.7 Processing and Estimation**

### **Criteria**

*22 Documentation for all data processes is maintained*

*23 Documentation for all systems, programs or applications is maintained*

*24 The processing system has been tested after the last revision*

*25 Raw data are saved in a secure location*

*26 The sampling bias and variance of the estimates are at acceptable levels*

**Processing** is the application of programs or procedures to a database for almost any purpose. For example, the data that arrive at CIHI are processed to test for errors, determine necessary edits, make modifications to the data and produce estimates.

**Estimation** is the aggregation of data, in any way, to produce a value that is used to represent the population of reference and to draw conclusions on the population. It is important to note that almost all values produced with databases (even aggregate totals) are estimates in the sense that they are approximations of reality, and not necessarily the true value.

The processing and estimation characteristic focuses on whether programs or systems affect the data quality. An error in an imputation program, or the improper inclusion or exclusion of data, can have a significant affect on the data quality of a database. It is important that all programs and systems related to the database be tested after each revision.



**Criterion 22** *Documentation for all data processes is maintained*

A process is the sequence of steps that is used when loading the data, editing the data, producing estimates, etc. It is important that database personnel know what to do, how to do it and what to do next when dealing with the data. Loss or relocation of staff can easily result in a loss of knowledge about the process, which can result in the process being followed incorrectly. The documentation should ideally be in one location, but a single location for each process is sufficient. The steps are sufficiently documented if a person new to the project could use the documentation to follow the processes.

This criterion is met if all the processes that are run by the database personnel are adequately documented.

**Criterion 23** *Documentation for all systems, programs or applications is maintained*

The systems, programs and applications used with a database can affect its quality. Therefore, it is important that they be documented. The reasons for documentation are simple: documentation allows easy validation of the programs and, if changes must be made, the documentation makes it easier to implement changes. Good documentation should be accessible and easily understood by someone new to the project.

This criterion is met if the systems, programs or applications that are used with the database are documented.

**Criterion 24** *The processing system has been tested after the last revision*

Although revisions are occasionally necessary to accommodate modifications to a database or registry (such as changes to data elements), changes to programs can have unexpected consequences. It is important to test not only the modifications that were made to the program to see if they have the expected results, but also the downstream effects of the changes. For example, changing the format of a data element from numeric to alphabetic character can affect programs that later treat the data element as numeric. Unit, system and user acceptance testing should be performed.

This criterion is met if the systems are tested when changes are made.

**Criterion 25** *Raw data are saved in a secure location*

Due to the need for verification and the fact that errors may occur in processing, it is important that the raw data, as it is provided by the data suppliers, is saved in a secure location. The data should be saved in such a way that they cannot be modified or deleted by accident. Having the unmodified data allows database personnel to return to the original data if errors have occurred during data processing, if disputes arise about the data or if the results of an analysis are questionable. If the raw data cannot be saved, the data database or registry should, at a minimum, be able to recover data from the previous stages of data processing.

This criterion is met if the data that arrives from data providers are saved in a secure location or if any changes to data made in the last three months can be undone.

**Criterion 26** *The sampling bias and variance of the estimates are at acceptable levels*

This criterion applies only to estimates that are based on a sample. Databases that do not use samples should rate this criterion as *not applicable*.

**Sampling bias** is the average of sampling errors on all possible samples. Sampling error is the difference between the estimate obtained from the sample and the true but unknown value of the parameter in the population.

The **variance** of an estimate is a measure of the variability of the estimates obtained when drawing all possible samples from the population of reference. The standard error or the coefficient of variation (cv) is often reported rather than the variance of the estimate. The standard error is the square root of the variance. The cv is used to obtain a measure of the relative variation of a distribution that divides the standard error by the estimate. It is often expressed as a percentage.

For databases that work on a sample basis, the variance or the standard error should be calculated for all estimates. Also, a methodologist should examine the degree of sampling bias (if any). If the sampling bias or variance of the estimate is found to be at a level that is not acceptable to the major database or registry clients, this criterion should be rated as *not met*.

This criterion is met if the sampling bias and the variance of the estimates are at levels that are acceptable to most users.

## 5. Timeliness Dimension

Timeliness refers primarily to how up to date the data are at the time of release. How current the data are is measured in terms of the gap between the end of the reference period to which the data pertain and the date on which the data become available to users. Timeliness is therefore closely associated with relevance, in that if this delay is too great, the data may no longer be relevant for the needs of users. Though data must be produced in time to assure relevance, acceptable timelines may vary across CIHI data holdings. More complicated databases or registries (that is, databases capturing longitudinal data), or those that are dependent on other databases for data, cannot be held to the same timelines.

If too much emphasis is placed on considerations of timeliness, accuracy may be compromised. For example, without sufficient time for clinicians to complete hospital charts, hospital discharge data might be timely but incomplete. It might be argued that this sacrifice in completeness is not worth the gain in timeliness. Sufficient time must also be set aside after the database or registry year-end close and prior to release in order to check the data and to document the limitations for users. As there is always more quality control and documentation that can be done, a balance must be struck between timeliness and accuracy. At a minimum, the recommended data quality documentation must be available in time for release.

The purpose of the timeliness dimension is to examine how current the data are and whether the recommended data quality documentation was made available in time for release. The criteria in this dimension also assess whether major database or registry reports are released on schedule. The dimension is comprised of the following characteristics:

- Data currency at the time of release—are the data made available in a reasonable amount of time?
- Documentation currency—are key documents released on time?

Dimension	Characteristics	Criteria
Timeliness	Data currency at the time of release	27 to 30
	Documentation currency	31 to 32

## 5.1 Data Currency at the Time of Release

### Criteria

- 27 *The difference between the actual date of release and the end of the reference period is reasonably brief*
- 28 *The official date of release was announced in advance of the release*
- 29 *The official date of release was met*
- 30 *Database or registry methods are regularly reviewed for efficiency*

This characteristic first and foremost helps determine how current, or up to date, the data within a database or registry are at the time of release. Data currency is the key component of timeliness and is measured by taking the difference between the date of release and the last date to which the data relate. The duration should be short enough so that the data remain relevant for their main purposes. Also pertinent to data currency is whether the data are released on time and whether the database or registry methods are as efficient as possible. If the methods used to process and analyze the data are as accurate and efficient as possible, the data will not be unnecessarily delayed.

**Criterion 27** *The difference between the actual date of release and the end of the reference period is reasonably brief*

The **date of release** is defined here as the official date upon which an annual subset of data from a database or registry becomes available to users. The **reference period** refers to the period of time which the data actually span or to which they relate. The start of the reference period is the first date to which the data relate, and the end of the reference period is the last date to which the data relate. For databases or registries that do not have an annual release of data, any major releases of data should be used as the points of comparison.

Different databases or registries will have different standards related to what is reasonably brief. As a general rule, a 12-month period between the end of the reference period and the release date is reasonable.

This criterion is met if the difference between the date of release and the end of the reference period is reasonably brief.

**Criterion 28** *The official date of release was announced in advance of the release*

Major releases, such as an annual subset of data, should have official release dates announced far in advance. This announcement should be made to the main users of the data, either internal or external. For example, at Statistics Canada, all major official releases are announced in advance in *The Daily*. This is an important consideration for users, because it enables them to in turn develop their own operational plans. Attainable dates of data availability should be set per release and, if these dates do not meet client needs, alternatives should be investigated.

This criterion is met if the official date of release for the annual subset of data was planned for and announced at least six months in advance.

**Criterion 29** *The official date of release was met*

It is important to users that an annual subset of data is released on time and as planned. The timing of the actual release date in relation to the planned release date may impact the production cycle of those who are dependent on the data. Monitoring the achievement of pre-announced release dates and changes to the release dates is recommended.

This criterion is met if the data were released on or before the official date of release for the annual subset of data.

**Criterion 30** *Database or registry methods are regularly reviewed for efficiency*

The programs or systems that are used to prepare and analyze the data should be reviewed in an ongoing manner to ensure that they are as efficient as possible. For example, multiple programs may be combined to reduce the amount of manual input and time required for data management, analysis and report creation. Comparing database or registry methods to similar external or internal databases may yield insights that result in improved efficiency. Existing methods may be reviewed in light of new technologies, procedures or standardized practices across databases that might be more efficient (such as electronic data capture or software) and may result in more accurate data at the same time. New methods or technologies may be an ideal way to improve timeliness and accuracy at the same time.

This criterion is met if database or registry methods are reviewed at least once in the last year.

## 5.2 Documentation Currency

**Criteria**

31 *The recommended data quality documentation was available at the time of data or report release*

32 *Major database or registry reports were released on schedule*

This characteristic guides in determining whether key documents were made available on time. More specifically, this characteristic is useful for knowing whether the recommended data quality documentation and any major database or registry reports were made available when needed or as planned. The purpose of data quality documentation is to inform users of the major limitations associated with the data, so that they can decide whether the data are fit for their intended use. This type of information is also necessary for the correct interpretation of results based on the data. It is therefore crucial that data quality documentation be made available, along with any major data release or report.

**Criterion 31** *The recommended data quality documentation was available at the time of data or report release*

It is important that data quality documentation be made available once users have access to the data or to reports based on the data. Therefore, a sufficient amount of time between database or registry close and release of the data must be allocated for data quality documentation, which is discussed in Chapter 9.

This criterion is met if data quality documentation was available at data or report release.

**Criterion 32** *Major database or registry reports were released on schedule*

It is important that the release dates for major database or registry reports be announced in advance and that the reports be released on schedule. Failure to meet deadlines with written reports not only inconveniences users, but can also undermine user confidence.

This criterion is met if the major database or registry reports were released on schedule.

## 6. Comparability Dimension

Comparability is defined as the extent to which databases are consistent over time and use standard conventions (such as data elements or reporting periods), making them similar to other databases. Within an organization like CIHI, with many different databases, comparability facilitates the understanding, interpretation and maintenance of the data. It is also directly related to the portion of CIHI's mandate that applies to the development and maintenance of a comprehensive and integrated health information system. Databases that are comparable will use the same data definitions, collect similar types of data and have the potential for record linkage with other similar databases. This in turn makes it possible to combine data from disparate sources in order to address important research questions that cannot otherwise be examined. Research on continuity of care is a prime example, given the range of clinical databases required for analysis (everything from emergency care to chronic care).

An additional advantage of comparability is that it can be used to assess other aspects of data quality, such as accuracy. Comparison of similar data systems can be an effective way of examining issues of coverage, coding errors and non-response.

The comparability dimension tells us how well databases meet a common standard. It is comprised of the following characteristics:

- Data Dictionary standards—does the database use CIHI standards for data definitions?
- Standardization—can common groupings be derived from the data?
- Linkage—can databases be joined by a common data element?
- Equivalency—are data values being converted correctly?
- Historical comparability—are data comparable over time?

Dimension	Characteristics	Criteria
Comparability	Data Dictionary standards	33 to 34
	Standardization	35 to 36
	Linkage	37 to 40
	Equivalency	41 to 42
	Historical comparability	43 to 45

## 6.1 Data Dictionary Standards

### Criteria

*33 Data elements are evaluated in comparison to the CIHI Data Dictionary*

*34 Data elements conform to the CIHI Data Dictionary*

This characteristic deals with the data elements in the database and how well they conform to the CIHI Data Dictionary, which contains the elements and definitions approved by the internal dictionary team. The goal is to have all databases use the same definitions for common data elements, thereby eliminating confusion among data submitters and researchers.

Data Dictionary standards are currently being reviewed and revised. While the adoption of the Data Dictionary standards is mandatory for newly developed databases or registries (or those being re-developed), the standards do not currently have to be applied to existing data holdings.

#### **Criterion 33** *Data elements are evaluated in comparison to the CIHI Data Dictionary*

The CIHI Data Dictionary is the standard for data elements that all databases at CIHI should follow. It is, however, in the process of development, so not all data elements are defined. Consequently, it is important that the existing databases be reviewed periodically against the CIHI Data Dictionary.

This criterion is met if the database has been evaluated against the CIHI Data Dictionary at least once in the past year.

#### **Criterion 34** *Data elements conform to the CIHI Data Dictionary*

Any data elements in a database that are also common to the CIHI Data Dictionary should share the same data attributes. There are several factors to consider when assessing conformance, including the data element name, the domain of values, the data type and the length. Ideally, everything should be the same. When it is not, conformance may be partial or not exist at all. For example, a data element representing the units by which age is measured should have the following attributes:

Name: AGE\_UNITS\_CODE  
Data Type: Alpha  
Maximum length: 3  
Domain: Y = years, M = months, D = days, N = newborn

A partial match occurs when only some of the attributes (such as length and domain) are the same and an exact match would occur if all the characteristics were the same. In general, differences in the values are more serious, since every record needs to be changed in order to conform to a given standard.



Note that the Data Dictionary is still a work in progress; therefore, assessing conformance only applies to those elements that are currently complete. Any justifiable deviations from CIHI standards should be described.

This criterion is met if at least 60% of the data elements common to the database or registry and the CIHI Data Dictionary are an exact match.

## 6.2 Standardization

### Criteria

*35 Data are captured at the finest level of detail as is practical*

*36 For any derived data element, the original data element is also maintained on the main database*

Databases often group elements in various ways, depending on the application or context. However, if it becomes necessary to compare data from different databases, a common grouping needs to be derived. Although it is neither practical nor reasonable to expect other databases to maintain the same groupings, capturing data at a sufficiently fine level of detail can ensure comparability. For example, if age of patient is typically reported in 10-year age categories, age in years should also still be available in order to derive other age groupings as needed. In this way, standardization of data elements across different databases can be achieved.

### **Criterion 35** *Data are captured at the finest level of detail as is practical*

A fine level of detail in data definitions is important, because it allows flexibility to conform to different standards. In most cases, the appropriate level of detail will be self-evident (full postal code, full diagnosis code and related prefixes, suffixes, etc.). Note, however, that this will differ depending on the database and, in some cases, the data supplier. For acute care stays, it is usually sufficient for length of stay to be measured in days. In contrast, wait time in emergency should be measured in minutes. Fine detail may not always be required for common uses, but it may be necessary in order to create new groupings.

This criterion is met if no more than two data elements are collected with insufficient detail. Any exceptions need to be justified.

### **Criterion 36** *For any derived data element, the original data element is also maintained on the main database*

As a general rule, data elements used in the creation of another data element need to be maintained with the database in the event that changes or new calculations have to be made. Sensitive items, such as health card number or birthdate, may require restricted access, but should never be completely deleted from the file. Simply maintaining the original element on the raw data file is not sufficient if accessibility is difficult. Note that this criterion applies to the original data and not to specific data requests that may require only the derived data elements.

This criterion is met only if no original data elements are permanently deleted from the main database.

## 6.3 Linkage

### Criteria

*37 Standard Geographical Classifications (SGC) can be used*

*38 Data are collected using a consistent time frame*

*39 Codes are used to uniquely identify institutions*

*40 Codes are used to uniquely identify persons*

**Linkage** refers to the process of joining records from two or more databases by the use of one or more common linking data elements, or joining records within a database or registry through a common data element. Given the variety of databases at CIHI, the capacity for linkage is extremely advantageous, as it allows one to combine data from different sources. Privacy and confidentiality guidelines must be adhered to when linking data holdings.

Ideally, linking data elements should share the same attributes, such as column name, width, type and format. At a minimum, the linking data element must be present on the database in order for linkage to be possible. In some cases, the linking data element will not be present because of the nature of the database. For example, expenditure databases do not collect patient-level data, so linking by patient is not possible on these databases.

This characteristic examines whether linkage is possible and not whether linkage is actually done. The criteria address the four main areas of linkage: geography, time, institution and person.

### **Criterion 37** *Standard Geographical Classifications (SGC) can be used*

This criterion relates to the **Standard Geographical Classifications (SGC)** as defined by Statistics Canada. Within this system, various geographical areas are grouped into a hierarchical system. The smallest types of aggregation include block face and enumeration area. These are nested within progressively larger groupings such as census tract and census division, culminating finally at province and country. Much social and demographic information derived from the census is aggregated at different levels of SGC, making it valuable for a wide range of research purposes. However, SGC is not often captured directly by databases. Instead, it is much more convenient to collect postal code, which can then be converted to SGC by way of the **Postal Code Conversion File**.

In some cases, geographic information can apply to more than one entity. Clinical databases, for example, should collect geographic information not only on the patient, but the facility as well.

This criterion is met if the entities on which data are collected (facilities, persons, province, etc.) are identifiable by either postal code (all six digits) or the relevant Standard Geographical Classification(s). If the lowest level of geography used is province, standard Canada Post province codes should be used.

**Criterion 38** *Data are collected using a consistent time frame*

Consistent time frames are important, not simply with respect to linkage, but also for making simple comparisons of summary data. It would be awkward, for example, to compare two estimates where one is based on calendar year and the other fiscal year.

This criterion is met if sufficient data are available that would allow the data to be arranged by either fiscal year or calendar year.

**Criterion 39** *Codes are used to uniquely identify institutions*

Institutions are a common level of analysis at CIHI and should therefore have a unique identification code in order to facilitate record linkage. Usually, the province assigns the institution identifier a numeric code. Other institution numbers are acceptable as long as a suitable cross-reference table is available, such as the Institutional Care Facility Master Inventory (ICFMI), used by Statistics Canada. Note that institution name by itself is not suitable for linkage purposes. Although names are invaluable as identifiers, they make poor linking data elements, given the inconsistencies in spelling, abbreviations and formatting that often can occur across different databases.

This criterion is met if an acceptable institution code (provincially assigned identifier or equivalent) is available on the database.

**Criterion 40** *Codes are used to uniquely identify persons*

The purpose of this criterion is to ensure that a suitable identifier is present that accurately distinguishes between persons in the database. In order to do this, the identifier must be unique, be consistent over time and have the capacity to accommodate future individuals. In order to facilitate record linkage, the data element must be consistent across databases as well. For clinical databases, this will most likely be the health card number. If a de-identified or encrypted data element is used, it should be possible to map the record back to the health card number. For personnel databases, other identifiers may be appropriate, such as those assigned by the province or regulatory body.

This criterion is met if a unique person identifier is available in the database that could be used to link to corresponding records in other databases.

## 6.4 Equivalency

### Criteria

*41 The impact of problems related to crosswalks and conversions falls into one of the predetermined categories*

*42 Methodology and limitations of crosswalks and conversions are documented*

Equivalency refers to how well data can be mapped from one format to another. Crosswalks and conversions are simply tables that are used to map one data format to another. Some examples relevant to CIHI include ICD-10-CA to ICD-9 or postal code to Standard Geographical Classification (SGC).

In the case of **conversions**, the mapping is one-to-one. For example, one could create a table that maps Canadian dollars to euros, or Celsius to Fahrenheit. In both of these cases, the conversion is simple, since the formulae are straightforward.

**Crosswalks** involve a many-to-one or one-to-many relationship. Assigning enumeration area (EA) to a single postal code is a good example of this, because a single postal code can map to more than one EA, or vice versa. Methodologies such as the Postal Code Conversion File have been designed to do this, but there may be some error involved. The success of a crosswalk or conversion is based largely on how well it can convert values from one format to another.

**Criterion 41** *The impact of problems related to crosswalks and conversions falls into one of the predetermined categories*

This criterion assesses the crosswalks and conversions used in the database. In general, any crosswalk or conversion should be thoroughly tested before it is implemented in a database. Misclassifications should be analyzed and adjustments made, if necessary. In addition to any first-hand experiences, one may want to consult relevant references in the literature about known issues. If the database uses more than one crosswalk or conversion, base the overall assessment on the weakest. Assess this criterion based on the following guidelines.

Suggested Rating	Guideline
Minimal	Little or no problems
Moderate	Identifiable problems that are limited in scope
Significant	A significant portion of codes are not being converted properly and this has an impact on results
Unknown	Equivalency has not been investigated

**Criterion 42** *Methodology and limitations of crosswalks or conversions are documented*

Due to the complexity of many crosswalks or conversions, the methodology and limitations need to be adequately documented. Any enhancements or alterations should be explained. Information derived by crosswalks or conversions also needs to be documented in reports as such. Simple crosswalks or conversions, such as single-year to five-year age groups need not be documented.

This criterion is met if crosswalks or conversions are adequately documented annually.

## 6.5 Historical Comparability

### Criteria

*43 Trend analysis is used to examine changes in core data elements over time*

*44 The extent of problems in comparing data over time falls into one of the predetermined categories*

*45 Accessible documentation on historical changes to the database exists*

Historical comparability refers to the consistency of data concepts and methods over time, which in turn allows one to make valid comparisons of different estimates at different points in time. Many things can make the comparison of data over time difficult. For example, database enhancements that will improve a database for the future can sometimes inhibit historical comparability. In such cases, one may have to introduce changes or enhancements that preserve the historical data or maintain careful documentation to alert the user of limitations.

**Criterion 43** *Trend analysis is used to examine changes in core data elements over time*

Trend analysis is defined broadly here. It includes comparisons of counts or proportions over time, as well as more sophisticated time series analysis, smoothing or curve fitting. Graphing data is sufficient and is often particularly helpful for investigating temporal changes. Within a clinical database, one might examine changes in the number of admissions for a particular disease group over the past several years. One of the primary rationales for longitudinal analysis is to detect any potential problems in the data as a result of changes in concepts or methodologies. Note that no change across years may also be an indication of a problem if the data are expected to naturally trend upward or downward.

This criterion is met if trend analysis is performed for core data elements within the past year.

**Criterion 44** *The extent of problems in comparing data over time falls into one of the predetermined categories*

It is important to take into account difficulties involved in producing valid trend estimates. Changes in methodology, inclusion criteria or unit non-response may make it impossible to determine whether the observed changes were real or not. For example, calculating the

total number of admissions from a particular acute care institution may be misleading if mergers or changes in institution type are not accounted for. When determining the number of physicians working in a province, a change in the inclusion criteria, based on the total amount billed to the province, may make past estimates invalid. The following is a rough guide for assessment of this criterion.

Suggested Rating	Guideline
Minimal	Little or no problems in producing comparable trends
Moderate	Problems have been identified with some trend data
Significant	Accurate trend data cannot be produced for a core data element
Unknown	Unknown whether accurate trends can be produced

#### **Criterion 45** *Accessible documentation on historical changes to the database exists*

This criterion assesses whether documentation of historical changes exists and is maintained in one document. It should include changes to concepts, methodologies, frame and data elements. Note that a set of manuals, each of which describes the current year changes, is not an acceptable form of historical documentation, as it becomes too difficult to track changes. Major changes from previous years should be included in the external data quality documentation, but a more detailed document for internal use may also be necessary.

This criterion is met if a single document of historical changes is maintained.

## 7. Usability Dimension

Usability reflects the ease with which a database's or registry's data may be understood and accessed. If data or other information products are difficult to use, they can be rendered worthless no matter how accurate, timely, comparable, or relevant they may be.

Several factors contribute to the usability of a database's or registry's data. In general, the greater the number of limitations or exceptions associated with the data, the more difficult the data will be to interpret. Efforts made to reduce or prevent data quality limitations or to improve the standardization of data improve not only the ease with which the data can be used, but also the accuracy of the data. Inconsistent database or registry methods may also complicate interpretation. The benefits derived from the introduction of new methods (for example, data element name or definition changes) should therefore be weighed against any loss in interpretability. Simply put, the fewer the limitations and changes, the easier the data will be to interpret.

To aid in the interpretation of the data, key users should be informed of any known major limitations at year-end and on an ongoing basis after release. Once major limitations are known, they should be documented for users. Database or registry methods and changes to the methods should also be documented for users. Also, the data have to be in a readily accessible user-friendly form. Finally, no matter how well documented or accessible, if users are not aware of a database's or registry's existence, the data will not be used.

The purpose of the usability dimension is to identify problematic aspects of a database or registry that are related to the interpretability of its data, as well as to identify how well documented and accessible the data are. It is comprised of the following characteristics:

- Accessibility—how readily accessible are the data?
- Documentation—how well documented are the data?
- Interpretability—how easy is it to understand the data?

Dimension	Characteristics	Criteria
Usability	Accessibility	46 to 48
	Documentation	49 to 51
	Interpretability	52 to 53

## 7.1 Accessibility

### Criteria

- 46 *An official subset of microdata is defined, created, made available and frozen per release for users, where appropriate*
- 47 *Standard tables and analyses are produced per release*
- 48 *Products are defined, catalogued and/or publicized*

This characteristic deals with the ease with which a database's or registry's data can be obtained from CIHI. This includes the ease with which the existence of the database or registry can be ascertained, as well as the suitability of the format of the data. Data that users do not know about, cannot locate, or cannot bring into their own working environment will not be of use to them.

**Criterion 46** *An official subset of microdata is defined, created, made available and frozen, per release for users where appropriate*

The data that are used for analysis and the creation of reports should be saved in a secure location for future reference. It is often necessary to refer back to previous sets of data in order to run further analyses. Having one version of the data set used in the creation of a report will ensure that results based on any new analysis will be consistent with the previously released results. Note that the data can be provided in various formats, depending on what the users want.

This criterion is met if a microdata subset of the database or registry is frozen per release.

**Criterion 47** *Standard tables and analyses are produced per release*

For many users, aggregate statistics or summary tables are more useful than microdata. In addition to major reports and/or microdata, aggregate statistics and standard tables should also be made available for users, per annual subset release. The standard tables are usually cross tabulations of core data elements on the database. The results based on the standard tables should be checked against those from previous years to confirm they are reasonable.

This criterion is met if commonly used standard tables and analyses are made available per annual release.

**Criterion 48** *Products are defined, catalogued and/or publicized*

To assist users, a database or registry and its associated products should be listed in the corporate-wide dissemination systems. These include the CIHI Web site and *Products and Services Catalogue*. The Web site can be used as a virtual library of all the information products that are available for the public from CIHI. Other channels, such as the press (via media releases) and public libraries, may also be used.

This criterion is met if a database or registry is listed in any of the CIHI dissemination systems per annual release.



## 7.2 Documentation

### Criteria

*49 Data quality documentation for users exists per annual subset release*

*50 Database or registry methods documentation exists for internal purposes per annual subset release*

*51 A caveat accompanies any official preliminary release*

This characteristic is helpful for knowing whether the documentation needed to understand the data is available. Documentation is necessary for appropriate interpretation and utilization of a database's or registry's data. Documentation normally includes a description of the underlying concepts, data elements, classifications used, methodology of data collection and processing, as well as information on the accuracy of the data. What has been measured, how it was measured and how well it was measured needs to be clearly documented for users.

#### **Criterion 49** *Data quality documentation for users exists per annual subset release*

The purpose of data quality documentation for users is to give external users of the database or registry sufficient information so they can decide if the quality of the data is appropriate for their intended use. It is primarily designed to outline the methods used in the collection and manipulation of the data and to provide the major limitations of the data. Contact information should also be provided with any release so that users can access additional information on the limitations they may require for their intended use.

A stand-alone data quality document for users, or its recommended equivalent, should be made available at least once a year (that is, per annual subset release).

This criterion is met if data quality documentation for users exists per annual subset release.

#### **Criterion 50** *Database or registry methods documentation exists for internal purposes per annual subset release*

To facilitate the interpretation and proper use of the data, all database or registry methods documentation should be made readily available to CIHI personnel who work with the data. This documentation is referred to as the "methods document" in section 3. For example, the processing methods and changes in processing methods could be made available electronically. While data quality documentation provides some background information and outlines the major data limitations, detailed background notes, as well as all known data limitations, should be made available internally.

This criterion is met if database or registry methods documentation exists for internal purposes per annual subset release.

### Criterion 51 *A caveat accompanies any official preliminary release*

In an effort to improve timeliness, some databases or registries may provide preliminary releases of data or results. Preliminary releases can be unofficial or official. Data quality documentation needs only accompany official preliminary releases.

An **unofficial preliminary release** is defined as any preliminary release of annual subset data that is designed to help certify or validate data. For example, prior to publication, health indicator counts might be sent to the health regions for verification. An **official preliminary release** is defined as a release of possibly incomplete annual subset data for the purpose of improved timeliness. For example, data that are collected on an annual basis might be released six months prior to year-end so that health care system planners gain an early indication of the complete data to follow.

For official preliminary releases, a caveat must be provided that advises that the data may not be complete and that they are subject to revision. A description of the response rate to date, as well as the expected final response rate, should be included. The anticipated revisions and their possible impact should also be conveyed.

This criterion is met if a caveat accompanies any official preliminary release of annual subset data.

## 7.3 Interpretability

### Criteria

*52 A mechanism is in place whereby key users can provide feedback to, and receive notice from, the product area*

*53 Revision guidelines are available and applied per annual subset release*

Interpretability refers to the ease with which the user may understand the data. Design features and underlying data quality limitations associated with data will largely determine its interpretability. For example, not only will an intricate population of reference limit the generalizability of the data, but it may also limit the ease with which the data can be understood. If standard concepts and classifications are in place, the data will be easier to understand and use.

Since the concept of interpretability is difficult to measure directly, this characteristic measures whether a mechanism is in place that facilitates interpretation and whether revision guidelines are in place.

**Criterion 52** *A mechanism is in place whereby key users can provide feedback to, and receive notice from, the product area*

Contact information should be included with releases so that users (internal or external) can share any major data quality limitations as they come to light. Similarly, there should be a mechanism that allows contact with major users, so they can be notified of the existence of any limitations that are discovered after the release. Information on

actions taken in light of the limitations and the effect of the errors should also be made available. Examples of such a mechanism might include a notification system for users or a users' group comprised of key users (for example, Statistics Canada, Health Canada and CIHI analysts).

This criterion is met if product area contact information is included with any major release and if major users are encouraged to use the contact information to provide feedback on any limitations they may discover.

**Criterion 53** *Revision guidelines are available and applied per annual subset release*

Initial estimates are often revised as errors or missing data come to light. A **revision** is a change to the data, or to the estimates based on the data, once the data have been placed in the public domain. If major limitations or updates are discovered after release, database-specific guidelines should be in place to aid in the decision of whether or not to release revised data. More specifically, the guidelines should state at what point the impact of newly discovered errors or updates would be severe enough to justify the release of a revised subset of data. The revision guidelines should also cover how and when revisions will normally be published. For example, a revision guideline might state that if national estimates are significantly affected by a post-release correction, then a revised data set is released.

This criterion is met if database-specific revision guidelines are available and applied per annual subset release.

## 8. Relevance Dimension

Relevance reflects the degree to which a database or registry meets the current and potential needs of users. Maintaining relevance requires keeping in touch with key users and stakeholders. Relevance is concerned with whether the available data inform the issues most important to users. In addition to ensuring that its data are accurate, timely, comparable and usable, to fulfill its mandate CIHI must also make certain that its data holdings continuously reflect Canada's most important health care information needs. The challenge is to balance the differing needs of current and potential users to produce a program that goes as far as possible in satisfying key needs.

The purpose of the relevance dimension is to assess how well a database or registry can adapt to change and whether the database or registry is perceived to be valuable. It is comprised of the following characteristics:

- Adaptability—can user needs be anticipated and planned for?
- Value—how valuable are the data?

Dimension	Characteristics	Criteria
Relevance	Adaptability	54 to 55
	Value	56 to 58

### 8.1 Adaptability

Criteria
<i>54 Mechanisms are in place to keep clients and stakeholders informed of developments in the field</i>
<i>55 The database or registry can adapt to change</i>

The adaptability of a database or registry relates to whether it is well positioned and flexible enough to address the current and future information needs of its main users. In order to remain relevant, a database or registry may have to adapt in an ongoing manner to emerging issues in the field. As needs and priorities change constantly, feedback mechanisms should serve to maintain awareness of the current and future issues of interest for each major client and stakeholder group.

If existing or developing issues are known and tracked, then future information needs may be anticipated. It is important to remain proactive and not reactive to main user needs. Once anticipated, future information needs can be factored into the design of the database or registry. Although it is impossible to predict the future needs of users with complete accuracy, one can try to design databases and registries that allow for change.

**Criterion 54** *Mechanisms are in place to keep clients and stakeholders informed of developments in the field*

Maintaining a client and stakeholder liaison by the product area staff serves to keep CIHI staff abreast of the current and emerging issues and of the information needs that are likely to result from the issues. Product area staff might keep in touch with main clients or stakeholders by arranging or taking part in expert group meetings, steering committees and professional advisory committees. Conference participation and the submission of papers to peer-reviewed journals may also be informative. If main users or stakeholders are common across holdings, then a coordinated approach to the consultative process may be considered across databases or registries.

This criterion is met if liaison mechanisms are in place to help clients and stakeholders stay abreast of developments in the field.

**Criterion 55** *The database or registry can adapt to change*

In order for a database or registry to remain relevant, ongoing changes may be necessary. In order to address emerging issues, existing data elements might need to be redefined, or new data elements might be added. For example, new date and time data elements may be added to capture emergency room (ER) wait time in an ambulatory care database. A database or registry should also be able to incorporate new technical standards as they arise—for example, the International Classification of Disease, 10th edition (ICD-10).

In addition to dealing with important emerging issues or with new technical standards, changes to a database or registry may also be required to deal with data quality limitations. For example, if negative lengths of stay are detected, new edits may be added. Adapting to emerging issues, incorporating new standards and dealing with data quality issues within a product area's control might all be considered part of ongoing database and registry improvement.

While flexibility in a database or registry is important, the benefits of any changes should be weighed against the potential loss in comparability or interpretability.

This criterion is met if a database or registry has demonstrated the ability to adapt to an important emerging issue, to a new technical standard, or to a major data quality limitation.

## 8.2 Value

**Criteria**

*56 The mandate of the data holding fills a health information gap*

*57 The level of usage of the data holding is monitored*

*58 User satisfaction is periodically solicited*

The value of a database or registry may be defined by its contribution to health or health care system knowledge and to its use. That is, the worth or utility of a database or registry depends on whether it fills a health or health care system information gap and whether it successfully serves to address its purpose.

In addition to keeping in touch with main users and stakeholders to maintain awareness of emerging information needs, the perceived value of a database or registry's data should also be monitored. To keep response burden low, the liaison mechanisms described previously (criterion 54) should be used to generate feedback on current programs in addition to information about future needs.

**Criterion 56** *The mandate of the data holding fills a health information gap*

The value of a database or registry, to a certain extent, depends on whether it fills a health information gap. The mandate of the database or registry should be periodically assessed in relation to the other data holdings within CIHI and across the field externally. How a database or registry complements the other CIHI holdings and how it compares to similar data sources in the field should be well understood.

This criterion is met if the mandate of the data holding fills a health care information gap.

**Criterion 57** *The level of usage of the data holding is monitored*

The value of a database or registry may be related to the extent that the data are used. Evidence of usage may include high-profile uses of the data (for example, the Romanow Report), Web site hits, press clippings, news items, citations, staff-authored papers, sales, appearances by staff in the media, conferences and policy forums.

This criterion is met if the level of usage of the data holding is monitored

**Criterion 58** *User satisfaction is periodically solicited*

It is important to assess whether the database or registry is satisfying user needs and to apply the results from the assessment in a program review. Stakeholder satisfaction survey results may be used as direct evidence of the perceived value of a database or registry. A satisfaction survey might also be an opportune time to solicit feedback on the perceived accuracy, timeliness, comparability and usability of the data. Internal analysts are a key source of feedback, and findings from any applicable internal review should be carefully considered, including a review of frequently asked questions (FAQ) if a client support hotline is in place.

This criterion is met if client satisfaction assessments are conducted at least once every four years.

## 9. Documentation

Documentation is an essential part of any database or registry. Documentation provides proper context for information, easy confirmation of facts and, among other things, an efficient method of information dissemination and storage. Although many types of documentation are necessary for a database, this chapter will focus on documentation related to data quality.

Although all documentation could be argued to affect data quality, the three main types of documentation that will be referred to here are:

1. *Data Quality Assessment Report*—an internal CIHI report that summarizes the results of the data quality assessment (Chapters 3 to 8);
2. Data quality documentation for users—documentation that is provided to users of the database or registry; and
3. Methods documentation—detailed documentation of the methodology of the database or registry.

### 9.1 The Data Quality Assessment Report

The data quality assessment is primarily intended for the use of the database or registry staff. The results of the assessments are internal to CIHI and should not be released to external users. However, other personnel at CIHI may use the results of the assessment for their own purpose, and it is therefore important that the results are documented and justified. It is important to remember that the purpose of the assessment and the related report is to aid in the documentation and improvement of data quality and not to rate the database or the personnel working on the database.

Data Quality personnel are available to assist in the development of the report and to review the final version. The manager of a database or registry is responsible for signing off on the completed assessment report.

The evaluation should be based on a recent subset of data. This may be the most recent year, if data are collected on a yearly basis, or a recent version of the entire database. The ratings should not reflect changes or improvements that will be made in the future, although they may be noted in the text.

The evaluation tool is comprised of 5 dimensions, 19 characteristics and 58 criteria.

Every **criterion** needs to be assessed as **met**, **not met**, **unknown** or **not applicable** and substantiated with a brief textual description and/or references to relevant documents or publications. Even criteria rated as *not applicable* need a line or two explaining *why* they are not applicable.

### **9.1.1 Suggested Format**

Since all assessment reports will be based on the data quality assessment, it is suggested that the reports follow the format laid out in the data quality assessment tool.

It is also suggested that the report start with a brief introduction describing the database or registry and the timeframe for the assessment.

Following this, the dimensions, characteristics and criteria should be reported on in the order they appear in the data quality assessment.

Recommendations are a key component of the evaluation process. Include them in relation to problematic criteria or characteristics within the body of the report. They should also be summarized in the executive summary.

An electronic version of the sample report template is available on CIHI's Data Quality intranet site.

## **9.2 Data Quality Documentation for Users**

The purpose of data quality documentation is to give users of the data sufficient information to decide if the quality of the presented information fits the intended use. It is important to present data quality issues in the context of what they mean to the user and how they can affect a user's analysis. The given information can also appear elsewhere: for example, the concepts and definitions can be presented within an overview of the database at the beginning of the attached publication. However, it is recommended that, for some databases, a complete stand-alone data quality document, with all the pertinent information that a user needs, be created. Any database that has a significant number of releases (either data releases or reports) should consider having a stand-alone data quality document to ensure that the limitations of the data are consistently highlighted, are easily accessible and can be distributed separately.

### **9.2.1 Suggested Format**

The data quality documents for a database should contain an introduction followed by seven main sections.

The introduction should provide the purpose of the CIHI data quality document (that is, to give users sufficient information to decide if the quality of the presented information fits the intended use.) The introduction should also include one to two paragraphs describing the data that are evaluated, the rationale for its existence, a summary of the major data limitations and references for more information regarding the data source. Finally, more detailed information regarding the time span covered by the data quality document, including specific dates of inclusion and/or exclusion, should appear in the introduction.

Following the introduction, it is recommended that the data quality document include, but not be limited to, the seven sections shown on the following pages.



## **1. Concepts and Definitions**

### **1.1 Mandate or Purpose**

The mandate or purpose of the database or registry should be given.

### **1.2 Population**

The population of the database or registry should be defined, typically clarifying the population of reference and adding a specific time period.

### **1.3 Data Elements and Concepts**

Core data elements and concepts should be defined. Exact formulae do not have to be given. It is not necessary to replicate the Data Dictionary; only the relevant information needs to be included.

## **2. Major Data Limitations**

The purpose of this section is to inform data users of the major data quality issues. Major data limitations, as well as their estimated impact or resolution, should be documented here, and any portion of the database that was identified as a significant data quality issue through the application of the data quality assessment tool should be discussed here.

While it is possible to include all the limitations of a data source, the data quality document for users may be restricted to the data quality issues and variables relevant to the major users. If a specific product or specific external user's request necessitates separate data quality documentation, only the relevant limitations need to be identified.

## **3. Coverage**

This section should include three parts: a description of the frame, a description of the frame maintenance procedures and a description of the impact of the frame maintenance procedures.

## **4. Collection and Non-Response**

Descriptions in this section should be as brief as possible. There is a strong temptation to describe in detail all the work that has been done collecting and collating data—but that is not the purpose of this section. This section should not only describe the collection procedures, etc. used, but should also explain how they may affect the external user's analysis. It should include a discussion of data collection, the quality control procedures used, applicable response rates, any adjustments for non-response (or lack thereof) and any known problems with bias or reliability.

## **5. Major Methodological Changes From Previous Years**

This section highlights where the current database or registry has changed from previous iterations (for example, cases where change in numbers may be due to change in procedures and may not reflect actual change). This section should include any major changes made to the database in the previous year and any planned future changes. If longitudinal comparisons are made, any relevant historical changes should be documented and references to relevant documentation included.

## **6. Revision History**

This section highlights changes or corrections made to data that had previously been presented (historical data). If the data or estimates used in previous years were changed in any way, users should be informed so that they are not confused when the revised numbers do not match the previously released data. The estimated impact or resolution of each major issue should be given.

## **7. External Comparability**

The last section is concerned with how comparable the data or estimates are to other sources. If the estimates or data are not comparable to similar estimates or data that have previously been released or are going to be released by sources external to the database, any differences should be noted.

### **Example**

Please see the Data Quality section's intranet site for an example of a data quality document for users.

## **9.3 Methods Documentation**

For efficient documentation of methodology, all relevant information should be drawn from one well-maintained and exhaustive source. CIHI is currently in the process of developing a standard for this documentation. The methods documentation should contain all detailed information about the database and the process that is used to collect and process the data.

At present, data flow and methodological information for each CIHI database may not be available from one source. For every CIHI database or registry, a complete document that outlines the entire data trail (including all methodological information) should exist and be maintained on an ongoing basis. The methods document should be the reference that is used for questions about the methodology of the database.

The purpose of the methods document is to document the data flow of a database or registry to the point that a person new to the project could take over the project if there were loss of staff. It is not always possible to have new staff trained by leaving staff, so there should be enough documentation about the processes used so that new staff can take over.

The methods document should be independent of the data received from data suppliers, as it focuses on the methodology used with the data and not with the data itself. The document should contain information on the steps for data processing, who is responsible for them, who should sign off on the steps and what data quality checks are in place. The documentation of policies and procedures used with a database or registry allows for a clear path to process improvement, audit trails, specification of responsibilities and a more reliable process. Good documentation allows the database or registry to become more process dependent than person dependent, which can result in increases in data quality and stability.

## Appendix A—Glossary

**Abstract:** A summary of information taken from a clinical chart. The process of summarizing the data is termed *abstraction*.

**Accessibility:** A characteristic. The ease with which a database or registry's data can be obtained from CIHI.

**Accuracy:** A dimension. How well information in the database or registry, or derived from the database or registry, reflects the reality that it was designed to measure.

**Adaptability:** A characteristic. The degree to which a database or registry is well-positioned and flexible enough to address the current and future information needs of its main users.

**Administrative database:** A database containing information that is primarily collected for record keeping, finances or purposes other than research.

**Aggregate statistics:** Both statistics on a large grouping (or aggregate) of data, such as provincial estimates, or statistics used to summarize (or to aggregate) the data, such as the mean or median.

**Annual subset release:** The standard set of data that is provided on a yearly basis to those that use the data.

**Assessment tool:** The core component of the data quality framework. It is comprised of 58 different criteria that are used to identify aspects of concern with relation to data quality and to assess the limitations and strengths of a database or registry.

**Bias:** An assessment to what degree systematic differences occur between reported values in a database or registry versus the values that *should* have been reported. This provides an indication of whether or not errors that are present occurred on a random basis. See also *correlated bias*.

**Canadian Conceptual Health Data Model (CHDM):** The CHDM, a product of the CIHI Partnership for Information Standards, is a generalized model of key subject areas that represent the domain of health care from a Canadian perspective. Along with providing an example of the scope of information applicable to health care, it shows how these subject areas are related to each other.

**Capture:** See *data capture*.

**Characteristic:** An aspect of data quality that is comprised of one criterion or more.

**CIHI Data Dictionary:** The Data Dictionary contains the elements and definitions approved by the internal dictionary team. The elements have been named in order to comply with the Canadian Conceptual Health Data Model and, where possible, to comply with international standards, such as HL7 and ISO.

**Coefficient of variation:** A statistical calculation used to obtain a measure of the relative variation of a distribution that divides the standard deviation by the estimate. It is often expressed as a percentage.

**Collection:** See *data collection*.

**Comparability:** A dimension. The extent to which databases are consistent over time and use standard conventions, such as standard reporting periods or data elements, that makes them similar to other databases.

**Consistency:** A measure of the variation of responses over repeated measurements. It is also referred to as reliability.

**Consistency edits:** Edits that are performed in combination across data elements to ensure consistency (for example, a man having a caesarean section would be a case of inconsistent data).

**Control tables:** See *standard tables*.

**Conversion table:** A table that uses one-to-one mapping to convert one data format to another. See also *crosswalk table*.

**Core data element:** Any data element in the database or registry that is routinely used in any analysis.

**Correlated bias:** A systematic error in a data element associated with another data element in the database.

**Coverage:** A characteristic. The degree to which the frame of a database describes the population of reference.

**Criterion (pl.: *criteria*):** A specific statement that relates to a detailed element of data quality. Each criterion is given a rating of either *met*, *not met*, *unknown* or *not applicable*. In select cases, criteria are rated according to other predetermined categories, such as *minimal or none*, *moderate*, *significant* or *unknown*.

**Crosswalk table:** A table that uses many-to-one or one-to-many mapping to convert one data format to another. See also *conversion table*.

**Curve fitting:** The process of fitting a curvilinear function (or curve) to data points. A curvilinear function is one whose value, when plotted, will follow a continuous (but not necessarily straight) line—such as a polynomial, logistic, exponential or sinusoidal curve.

**Data attributes:** The characteristics of a data element. For example, the data element name, the domain of values, data type and width.

**Data capture:** The entering of data into a usable format by the data provider. Data capture may be done in a manual or electronic format.

**Data collection:** The gathering of supplied data from different data providers into a common database or registry.

**Data currency:** A characteristic. Data currency is the key component of timeliness and is measured by taking the difference between the date of release and the last date to which the data relate.

**Data Dictionary standards:** A characteristic. See *CIHI Data Dictionary*.

**Data flow:** The path data takes as it is processed. Often described through the use of a diagram showing the various data sources, processing steps and outputs.

**Data Quality Assessment Report:** A CIHI internal report that summarizes the results of the Data Quality Assessment.

**Data quality documentation for users:** Documents providing sufficient information so that data users can decide if the quality of the data is appropriate for their intended use. It is primarily designed to outline the methods used in the collection and manipulation of the data and to provide the major limitations of the data.

**Data quality work cycle:** A three-component approach to data quality, which involves a set of planning, implementing and assessing activities.

**Data provider:** The person or organization that provides the data to a database or registry. Data providers normally perform the function of data capture. They are also sometimes referred to as data suppliers.

**Data submission form:** A paper or electronic template that sets out the format of data elements for submission.

**Data supplier:** See *data provider*.

**Data type:** The format of a data element. For example, numeric, character or date format.

**Date of release:** The official date upon which an annual subset of data from a database or registry becomes available to users.

**De-identified:** Data that has been stripped of information that could be used to identify an individual or institution.

**Derived data element:** A data element that is a composite of other data elements.

**Dimension:** The distinct components that encompass the broader definition of data quality.

**Documentation:** A characteristic. Information on database or registry data quality, methods and caveats.

**Documentation currency:** A characteristic. The criteria falling under this characteristic examine whether the recommended data quality documentation and any major database or registry reports were made available when needed or as planned.

**Domain of values:** The range of values permitted in a given data element.

**Dual capture:** The act of collecting the same data twice and comparing the two copies to minimize errors of data entry and submission.

**Edit rules:** Rules used to identify missing or incorrect values in a database or registry.

**Edit reports:** A report that identifies if the records passed or failed data edits and why they failed.

**Editing:** The process of identifying missing or incorrect data in a database or registry.

**Equivalency:** Characteristic. How well data can be mapped from one format or standard to another. Crosswalk or conversion tables are often used for this purpose. An example would be the equivalency of ICD-10-CA codes to earlier ICD-9 codes.

**Estimation:** The aggregation of data to produce a value that is used to represent the population of reference and to draw conclusions on the population.

**Expert group:** A panel of key database contacts, both internal and external to the organization, created to advise on the maintenance and growth of a database.

**Flag:** A way of identifying a special case, often done through the creation of an additional data element.

**Frame:** A list of all units (for example, provinces, institutions or doctors) that is used to ensure that all units in the population of reference are collected. The frame of an administrative database can then be used to determine what proportion of the data was actually received.

**Frame maintenance:** A set of procedures to add any new units in the population of reference to the frame, as well as to remove any units that are no longer in the population of reference.

**Frame maintenance procedures:** Any practices or procedures that are used to update the frame.

**Grouping:** A way of relating similar variables by placing them in categories.

**Historical comparability:** A characteristic. The consistency of data concepts and methods over time that allows one to make valid comparisons of estimates from different time periods.

**Health Level Seven (HL7):** Standards for the exchange, management and integration of data that support clinical patient care and the management, delivery and evaluation of health care services.

**ICD-9:** International Classification of Diseases, 9th revision. A set of internationally accepted codes for classification of medical diagnoses and conditions.

**ICD-10-CA:** Canadian enhanced version of the International Classification of Disease, 10th edition.

**Imputation:** The process of determining and assigning replacement values for incorrect or missing data that should not be blank identified at the editing stage.

**Institution response rate:** The percentage of institutions that submitted data out of all institutions on the frame.

**Interpretability:** A characteristic. Refers to the ease with which the user may understand the data.

**ISO:** The International Organization for Standardization.

**Item non-response:** A characteristic. Also referred to as partial non-response. Covers blank data elements found on a record received that should not be blank.

**Item non-response rate:** The item response rate is the percentage of data elements for which data were reported out of all reporting records that should have reported the data element.

**Item response rate:** The rate of data elements (values) that are missing in comparison to the number of records that *were* submitted, not the number that *should have* been submitted. To get the rate, multiply the number of data elements for which data was reported by 100 and divide by the number of reporting records that should have reported the data element.

**Kappa statistic:** Also referred to as the *kappa coefficient*. The kappa coefficient measures the pair-wise agreement among a set of coders that are making category judgments and corrects for expected chance agreement.

**Linkage:** A characteristic. How easily the database or registry can be linked to other databases or registries.

**Linking data element:** A data element that is common to two or more data sets and may be used to combine them.

**Longitudinal data:** Data that span a length of time and thus may be used to see changes over time (for example, 10 years of physician salary information).

**Manual imputation:** The process of determining and assigning replacement values for the incorrect or missing data that should not be blank by hand rather than through an automatic process.

**Master methods document:** A well-maintained and exhaustive source that contains all of the detailed information about a database or registry and the procedures that are used to collect and process the data.

**Measurement error:** A characteristic. The difference between the value that is reported in the database or registry and the true but unknown value that should have been reported.

**Met:** A rating. Signifies that the requirements for the criteria have been achieved.

**Methods:** The statistical procedures used in processing and analyzing data, or more generally, the processes used in the running of a database.

**Microdata:** The data that are used for analysis and the creation of reports.

**Minimal:** A rating. Signifies that the level being measured is a low one.

**Missing at random:** Data that are absent from a database or registry on a random basis (the fact that they are missing is not related to any other data element).

**Moderate:** A rating. Signifies that the level being measured is a medium one.

**None:** A rating. Signifies that there are no occurrences of the attribute being measured.

**Non-response:** Failure to obtain data on all the units on the sampling frame. See also *unit non-response* and *item non-response*.

**Non-subjective variable:** A data element with a value that is not easily influenced by personal beliefs or feelings, such as a birthdate.

**Not applicable:** A rating. Signifies that the requirements for the criteria cannot be achieved.

**Not met:** A rating. Signifies that the requirements for the criteria have not been achieved.

**Official preliminary release:** The release of a possibly incomplete subset of data for the purpose of improved timeliness. For example, data that are collected on an annual basis might be released six months prior to year-end so that health care system planners can get an early indication of the complete data that will follow.

**Operationalize:** To convert to a form that may be worked with.

**Out-of-scope records:** Records that should not be included in the database or registry. This includes receiving records from units that are outside of the population of reference. See also *over-coverage*.



**Overall error:** An assessment of the degree that values that are reported in the database or registry match the values that should have been reported, the true values. This is similar to a validity check in epidemiological terms and provides an indication of the number of times that a data element is coded correctly.

**Over-coverage:** The situation where units that are not part of the population of reference are included on the frame, when duplicate records appear in the database or when out-of-scope records are included. See also *out-of-scope records*.

**Population of interest:** The population for which information is wanted in a statistical study. In many cases, information for the complete population of interest is not available. For example, a population of interest may be all hospitals in Canada with at least one acute-care bed.

**Population of reference:** The available population for which statements are made in a statistical study. For example, the population of reference may be all publicly funded, non-prison hospitals with at least one acute-care bed in all provinces and territories that were open for business on January 1 of the reference year. Ideally, the population of reference will be as close to the population of interest as possible in any study.

**Postal Code Conversion File:** A file that links postal codes to enumeration areas.

**Processing:** The application of programs or a sequence of procedures to a database. Processing can be done for many reasons, including producing estimates or testing for errors.

**Product area:** A department or group at CIHI that works to produce a certain database or deliverable.

**Production cycle:** The processes used to produce a regular set of reports or deliverables from data collection through processing and creation of the final outputs.

**Program:** A set of electronic instructions for data processing.

**Rate of over-coverage:** A rate calculated from the number of units, on the frame but not in the population of reference, multiplied by 100, divided by the number of units in the population of reference.

**Rate of under-coverage:** A rate calculated from the number of units, not on the frame but in the population of reference, multiplied by 100 and then divided by the number of units in the population of reference.

**Raw data:** The data that arrive from data providers and are not modified.

**Re-abstraction studies:** A study where designated experts go through the same data collection process normally done by the data supplier. The results of the designated experts are then compared with the results of the data providers.

**Record linkage:** The process of joining records from two or more databases by the use of one or more common linking data elements. Ideally, linking data elements should share the same attributes (such as data element name, width, type or format).

**Reference period:** The time period for the related data. The start of the reference period is the first date to which the data relate and the end of the reference period is the last date to which the data relate.

**Release:** Any report, data release or output from the database or registry.

**Relevance:** A dimension. The degree to which a database or registry meets the current and potential future needs of users. Relevance is concerned with whether the available data will inform on the issues most important to users.

**Reliability:** See *consistency*.

**Response burden:** A measure of how difficult it is for data suppliers to provide information. In many cases, response burden is quantified in terms of how long it takes for suppliers to gather and input information.

**Revision:** A change to the data, or to the estimates based on the data, once the data have been placed in the public domain.

**Revision guidelines:** Database-specific guidelines to aid in the decision of whether or not to release revised data. More specifically, the guidelines should state at what point the impact of newly discovered errors or updates would be severe enough to justify the release of a revised subset of data.

**Revision history:** A description of changes or corrections made to data that had previously been presented (historical data).

**Sampling bias:** The average of sampling errors on all possible samples. This differs from bias in measurement error (see *bias*), which results from a bias in the records.

**Significant:** A rating for criteria signifying that the level being measured is quite high. Also refers to something being important or statistically significant.

**Smoothing:** Smoothing techniques are statistical techniques used to reduce irregularities (random fluctuations) in time series data. They provide a clearer view of the true underlying behaviour of the series.

**Standard conventions:** A set of usual processes or attributes. For example, standard data elements or reporting periods.

**Standard error:** The square root of the variance.

**Standard Geographical Classifications (SGC):** A hierarchical classification system defined by Statistics Canada that groups various geographical areas. The smallest aggregate areas are enumeration areas, which are in turn nested progressively into larger groupings, such as census tracts and census divisions, culminating in provinces and countries. This classification is often accomplished using the Postal Code Conversion File to link from postal codes.

**Standard tables:** A set of tables that are produced every data cycle, against which results may be checked over time.

**Standardization:** A characteristic. An assessment of the level to which common groupings can be derived from the data.

**Statistical significance:** The likelihood that the findings were not produced by chance.

**Subjective variable:** A data element with a value that is easily influenced by personal beliefs or feelings, such as level of impairment.

**Temporal changes:** Changes over time.

**Timeliness:** A dimension. A measure of how current or up to date the data are at the time of release. How current the data are is measured in terms of the gap between the end of the reference period to which the data pertain and the date on which the data become available to users.

**Under-coverage:** A situation where a unit that should be part of the frame is not listed in it.

**Unit non-response:** A characteristic. Unit non-response refers to the units or entire records that belong to the frame *for which no information was submitted*. It is often confused with under-coverage, wherein information is missing on units that are not listed on the frame.

**Unit response rate:** A rate calculated from the number of units that submitted data to a database or registry, multiplied by 100, then divided by the number of units on the frame.

**Unknown:** A rating. Signifies that it cannot be determined whether the requirements for the criteria have been met or at what level they can be measured.

**Unofficial preliminary release:** Any preliminary release of a subset of data that is designed to help certify or validate data. For example, prior to publication, health indicator counts might be sent to the health regions for verification.

**Usability:** A dimension. A measure of the ease with which a database's or registry's data may be interpreted, understood and accessed. The issue of usability also relates to potential users knowing of a database's or registry's existence and the data being in a readily accessible, user-friendly form.

**Usage:** The extent to which the data from a database or registry are used.

**Validity checks:** Checks done to ensure that the proper format is used for a data element and that the response rate is appropriate. Validity checks can consist of comparing the response to a list of acceptable responses.

**Value:** Characteristic. The contribution of a database or registry to health or health care system knowledge and to its use.

**Variance:** A measure of the variability of the estimates obtained when drawing all possible samples from the population of reference.

**Width:** A data attribute referring to the number of characters permitted in a data element.

## Appendix B—Bibliography

Brackstone, Gordon. "Managing Data Quality in a Statistical Agency." *Survey Methodology* 25, 2 (Dec. 1999): pp. 139–149.

Statistics Canada. *Policy on Informing Users of Data Quality and Methodology*. (Policy 2.3.) Online, from <<http://www.statcan.ca/english/concepts/inform.htm>>.

Statistics Canada. *Statistics Canada's Quality Assurance Framework*. Ottawa: Statistics Canada, 2002. Catalogue no. 12-586-XIE.

Statistics Canada. *Statistics Canada Quality Guidelines*. (Third edition.) Ottawa: Statistics Canada, 1998. Catalogue no. 12-539-XIE.



