

accuracy  
timeliness  
comparability  
usability  
relevance

The CIHI  
Data Quality Framework

2009



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

Phone: 613-241-7860

Fax: 613-241-8120

[www.cihi.ca](http://www.cihi.ca)

ISBN 978-1-55465-696-7 (PDF)

© 2009 Canadian Institute for Health Information

How to cite this document:

Canadian Institute for Health Information, *The CIHI Data Quality Framework, 2009* (Ottawa, Ont.: CIHI, 2009).

Cette publication est aussi disponible en français sous le titre : *Le cadre de la qualité des données de l'ICIS, 2009.*

ISBN 978-1-55465-697-4 (PDF)

# The CIHI Data Quality Framework

2009

## Table of Contents

Acknowledgements .....	iii
1. Introduction .....	1
1.1 What Is CIHI’s Data Quality Framework? .....	1
1.2 What Is New in the 2009 Version of the Data Quality Framework? .....	2
1.3 What Is Data and Information Quality? .....	2
2. The Data Quality Work Cycle .....	3
3. Assessment of Data Quality .....	5
3.1 The Assessment Tool .....	5
3.2 Dimensions of Data Quality .....	6
3.3 Integrating Data Quality Dimensions .....	7
3.4 Ratings .....	7
3.5 Action Plan.....	8
4. Documentation for Data and Information Quality.....	9
4.1 The Data Quality Assessment Report .....	9
4.2 Data Quality Documentation for Users .....	11
4.3 Metadata Documentation .....	13
Appendix A—Significant Changes From 2005 Version.....	15
Appendix B—Summary of Dimensions, Characteristics and Criteria .....	17
Appendix C—Data Quality Framework Assessment Tool .....	21
Appendix D—Data Quality Assessment Report Template .....	79
Appendix E—Subcategories for Metadata Documentation .....	101
Appendix F—Glossary .....	111
Bibliography .....	121



## **Acknowledgements**

The Canadian Institute for Health Information (CIHI) acknowledges the contribution of many individuals at both the Canadian Institute for Health Information and at Statistics Canada to this and previous versions of *The CIHI Data Quality Framework*.



# 1. Introduction

Data and information quality are intrinsic to the Canadian Institute for Health Information's (CIHI) mandate to inform public policy, support health care management and build public awareness about the factors that affect health. CIHI engages in rigorous activities to ensure that the data collected and provided is of the highest standard.

CIHI has implemented a comprehensive program that includes processes and policies to continuously improve data and information quality, both within CIHI and in the broader health sector. Our corporate strategy, found in the following six-point plan, includes a number of initiatives aimed at prevention, early detection and resolution of data issues:

- Foster a data quality culture;
- Strengthen data quality infrastructure and capacity;
- Cultivate the data supply chain;
- Enhance external data quality collaboration;
- Promote communication and provide consultation; and
- Initiate the fast-track priority projects fund.

The complete CIHI data quality corporate strategy can be found at [www.cihi.ca](http://www.cihi.ca) in two documents entitled *Earning Trust* and *Earning Trust Three Years Later*.

Improving data and information quality is a collaborative effort. CIHI staff work jointly with data providers and CIHI data users in support of good data quality. This collaborative effort is required to meet the changing and expanding user requirements and expectations of CIHI data holdings. These increased requirements and expectations have prompted CIHI to conduct regular reviews of its Data Quality Framework.

## 1.1 What Is CIHI's Data Quality Framework?

This framework provides an objective approach to applying consistent data-flow processes that focus on data quality priorities, assessing the data quality of a data holding and producing standard data-holding documentation with the ultimate goal of continuous improvement in data quality for CIHI's data holdings. The framework encompasses three main components:

- A data quality work cycle
- A data quality assessment tool
- Documentation about data quality

**NEW**

## **1.2 What Is New in the 2009 Version of the Data Quality Framework?**

This 2009 version of the CIHI Data Quality Framework is an updated version of the 2005 CIHI Data Quality Framework. CIHI has experienced significant growth within the last number of years in the variety of data-holding types it has in the organization. With such growth it became quite evident that the framework would have to be enhanced from the 2005 version. This enhancement was mostly due to the uniqueness of data holdings in health personnel, drugs, health expenditures, medical equipment and home and continuing care with respect to clinical data holdings, which was largely the basis for previous data quality frameworks.

### **Challenges in Assessing CIHI Data Holdings**

Using one assessment tool to assess all of CIHI's data holdings is a challenge, given the diversity of these holdings. Some of the issues encountered include the following:

1. Multiple levels of data providers versus a single data provider;
2. Aggregate-level data versus record-level data;
3. Voluntary versus mandated submission;
4. Sample survey data versus census data; and
5. Longitudinal data versus point-in-time data.

This latest version of the framework addresses these challenges as comprehensively as possible. The details outlining the changes from the previous 2005 version can be found in Appendix A.

## **1.3 What Is Data and Information Quality?**

At CIHI, data quality is defined in the context of the users of our data. These customers include health system planners, ministries of health, regional authorities, data providers, health profession associations, researchers, health care providers and other special interest health organizations. If CIHI's data satisfies their particular data needs, then the data is said to be "fit for their use." To be fit for use, data must possess three attributes: utility, objectivity and integrity, according to the U.S. Census Bureau in 2006. Utility refers to the usefulness of the data or information for its intended users. Objectivity refers to whether the data or information is accurate, reliable and unbiased and is presented in an accurate, clear and unbiased manner. Finally, integrity refers to the security or protection of data or information from unauthorized access or revision. Decisions about the management of the health care system are made based on information provided from the data. Data is an asset for any organization. It is this data that is the basis for the production of information upon which decisions are made. The management of information quality in a secondary data-collector organization such as CIHI is paramount. High information quality is achieved by identifying information defect root causes, error-proofing the information processes, identifying information quality requirements and controlling information processes. The application of these sound management principles to information ensures that CIHI continues to be relied upon as a trusted custodian of quality pan-Canadian health data.



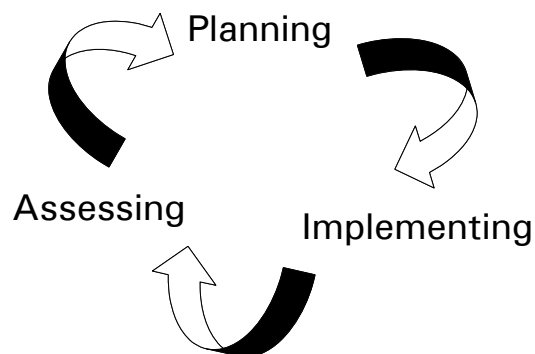
## 2. The Data Quality Work Cycle

Consistent data work-flow processes that focus on data quality priorities start with well-defined roles and responsibilities for data quality within an organization, and the implementation of a coherent work cycle for handling data and information.

At CIHI, data quality is a responsibility shared by each staff member, regardless of position or work department. Data and information quality are corporate priorities. Senior management provides overall direction for both data and information quality, sets priorities for corporate and ongoing data quality projects and ensures adequate resources for these initiatives. A data quality council, made up of managers from data holdings and users of their data, meets monthly and reports as required to the senior management team. Each program area responsible for a data holding has duties that include analyzing and evaluating their data and data processes to identify data quality issues, preparing plans to address the identified issues, finding ways and initiating practices to improve the data holding, conducting special studies and documenting their data quality for both internal and external users. A dedicated department has been established at CIHI to provide support to every CIHI data-holding program area and users of the data in the identification and resolution of data quality issues by developing tools such as the Data Quality Framework, providing methodological and statistical advice, conducting special studies, mining their data, producing corporate reporting tools such as the jurisdictional reports on data quality to the deputy ministers of health and engaging external stakeholders in the discussion on data quality and how to implement it across the health care system.

Specific details on the roles and responsibilities of CIHI's program areas and Data Quality department with respect to the implementation of this Data Quality Framework can be found in a document entitled *CIHI's Database Support Resource Guide*. CIHI staff can find this document on the Data Quality department's intranet page.

The **data quality work cycle** used at CIHI includes three types of activities: **planning** for a data quality activity (that is, a change to a set of edits applied to the data at source to ensure its quality prior to loading into the database); **implementing** this activity; and then finally **assessing** whether the desired objective was met (that is, were the edits correctly written and is only high-quality data loaded into the database). These three types of activities are repeated as often as needed to achieve the data quality objective.



**Planning** includes the activities necessary to prepare and prioritize the processes required for a data holding, as well as the design of any changes that are needed.

**Implementing** includes developing the processes needed and applying them to the data holding (such as collecting data, monitoring incoming records and releasing written reports). The results of implementation activities for one process can be useful in the planning of similar future processes.

**Assessing** involves evaluating the quality of the data holding 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. As mentioned earlier, the results of the assessment activity for a process can assist in future planning activities.

This three-stage work cycle is flexible and can be applied at any phase of a data holding's development and ongoing maintenance cycle. By focusing on integrating quality work processes within each activity of the work cycle, efficiency and a quality product will be achieved.

## 3. Assessment of Data Quality

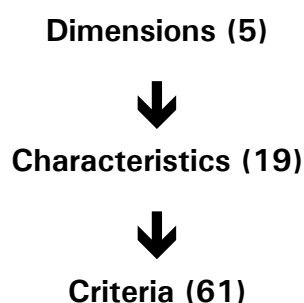
Assessing the data quality of each CIHI data holding is accomplished using the framework's **assessment tool**, the core component of the framework. The tool examines every aspect of the data quality of a data holding, but no tool can conduct an exhaustive assessment, as data quality is so broad. The assessment tool will identify strengths in the data and in the implemented data practices for the assessed data holding, as well as issues that need addressing. An action plan is required that identifies how and when these issues will be addressed.

The assessment tool examines five aspects of the data quality of a data holding. These five aspects are called "dimensions" in the tool and are accuracy, timeliness, comparability, usability and relevance. The definitions for these five dimensions are explored further in Section 3.2. Separating data quality into components allows the user to quickly identify specific aspects of the data that may create problems with regard to fitness for use. Although many organizations have organized data quality into very similar dimensions, there is no one way to describe these dimensions. CIHI's five dimensions encompass the same ones that are included in Statistics Canada's framework, although they may be labelled differently. Even though organizations may use different names for these dimensions, the need to explore different aspects of data quality is an accepted practice.

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

### 3.1 The Assessment Tool

The assessment tool comprises dimensions, characteristics and criteria. Each dimension is divided into related characteristics, and each characteristic is further made up of several criteria. These criteria are questions about each aspect of data quality. Once answered, it is this information that informs the program area about the presence of data quality issues.



CIHI's Data Quality Framework is based on 5 dimensions of data quality, 19 characteristics and 61 criteria. Each criterion is a question whose aim is to query a certain aspect of the data or data holding. Each criterion is generally either rated as *met* or *not met*. In a few instances, a criterion is rated based upon a range of values. It is the response to each of these questions that guides the data holding staff to the creation of the action plan for improvements described in Section 3.5.

A summary of this assessment tool is provided in Appendix B. The comprehensive Data Quality Framework Tool is included in Appendix C.

## 3.2 Dimensions of Data Quality

The dimensions are distinct components that encompass the broader definition of data quality. The five quality dimensions at CIHI are defined as follows:

### 1. Accuracy

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

### 2. Timeliness

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

### 3. 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 comparable to other databases.

### 4. Usability

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

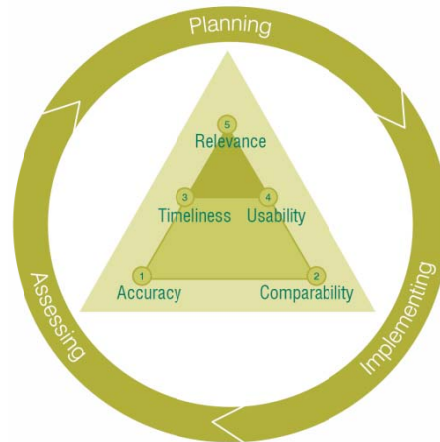
### 5. Relevance

Relevance reflects the degree to which a data holding meets the current and potential future needs of users.

**NEW**

### 3.3 Integrating Data Quality Dimensions

These dimensions can be overlapping and interrelated. Assessing fitness for use involves adequately managing each dimension. In addition, failure in one dimension might impair or undermine the usefulness of the final report or data release. It is important to find a good balance among the dimensions in terms of issues found. Action undertaken to address one dimension may positively affect other dimensions of quality. These five dimensions can be viewed together in the following hierarchical diagram within the data quality work cycle that was described in Section 2.



This hierarchical approach displays the relevance dimension as the most important dimension. If data is not relevant, its value decreases substantially even if the other four dimensions have been met. In order for the data to be relevant, it must meet both the current and potential future needs of the users. The building blocks for this hierarchical model require that the data be accurate and use consistent conventions to ensure its comparability to like data. Once these are established, data must be as current as possible so that decisions are made with recent information. Also, to ensure its usability, the data must be accessible and easy to understand. These four dimensions of data quality, if achieved, allow the user to make decisions that are based upon accurate, comparable, timely and usable data.

### 3.4 Ratings

Ratings are used in the assessment tool as a guide to highlight strengths and identify limitations of the data. Ratings are only a guide for program areas to consider in determining fitness for use. The subjective nature of data quality and the differing purpose of data holdings mean that no rating system will be able to identify all problems with data quality. High rating scores do not necessarily mean a data holding's data is problem free. Similarly, scores that indicate some data quality issues do not necessarily mean that a data holding's data should not be used. It is the responsibility of the program area to identify the presence or absence of data quality issues in its data holdings and to document the assessment results to inform users of any data quality issues.

In the majority of cases, each **criteria** in the assessment tool is given a rating of *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 rating table that can be used to determine what rating should be assigned. Program areas are responsible for determining the ratings for their data holdings. The ratings promote standard comparisons between databases and between data years within a particular data holding.

**It should be emphasized that rating scores are for internal purposes only and appear only in the assessment report. They are not to be included in any external documentation.**



### 3.5 Action Plan

Once the assessment of the data holding is completed, key findings from the assessment are summarized and an action plan is developed that identifies strategies to remedy any deficiencies. An example of an action plan that identifies recommendations, timeline for implementation and the person responsible is provided in Table A. Follow-up by program area managers is required to ensure that data quality improvements are made.

**Table A Action Plan Example**

Recommendation	When Initiated?	Staff Responsible	Target Date or Ongoing	DQ Involvement?
1. Investigate cause of duplicates contributing to over-coverage	2007–2008	Alan	2008–2009	Yes
2. Data elements need to be evaluated in comparison to CIHI Data Dictionary	2008–2009	Barbara	Sept. 2008	No
3. Validity checks performed on each data element and any invalid data flagged	2006–2007	Barbara Caroline	Yearly	No
4. Create master methods documentation	2007–2008	Alan Caroline Donald	2008–2009	Yes
5. User satisfaction needs to be assessed	2006–2007	Peter	2007–2008 Dec. 2008 (New Target)	No

## 4. Documentation for Data and Information Quality

An integral component of CIHI's Data Quality Framework is the **documentation of data and information quality**. Informing users about the quality of CIHI's data is a very important part of our data quality program, as this allows users to determine if the data is appropriate for their use. This information is applicable to both internal CIHI staff and external users of our data.

Documentation is an essential part of any data holding. 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.

Three types of documentation about a data holding's data quality are required by CIHI's Data Quality Framework:

1. *Data quality assessment report*—an internal CIHI report that summarizes the results of the data quality assessment;
2. *Data quality documentation for users*—documentation that is provided to users of the data holding; and
3. *Metadata documentation*—detailed documentation about the data holding.

### 4.1 The Data Quality Assessment Report

#### 4.1.1 Purpose

The data quality assessment report is primarily intended for use by the data holding staff. This internal report provides a record of the results of the assessment conducted to identify strengths of the data holding as well as any data quality issues. The data quality documentation for users is the document that is released for external publication, and its content is based largely on the internal report.

Refer to Section 4.1.4 for the suggested content and format when completing an assessment report for a data holding.

#### 4.1.2 When to Produce the Report

The evaluation should be based on a recent subset of data. This may be the most recent year, if data is collected on a yearly basis, or a recent version of the entire database. The ratings should reflect the current status of the data holding, although expected future changes or improvements can be noted in the text. It is possible to answer some of the criteria questions without having to wait for all of the data to be received. For other questions, all data will have to be received before the final assessment report can be completed. All criteria must be assessed before any major release of data or the publication of an annual report. The internal assessment report should be produced no later than three months after the release of the data or publication.

### 4.1.3 Review and Support by Data Quality Staff

Data Quality department staff are available to provide guidance during the criteria-assessment process and to assist in developing the report and reviewing the final version. Specifically, once the report is written by data holding staff, Data Quality department staff review the report and provide comments within the month. Once these comments are reviewed and/or incorporated, the manager of the data holding signs off on the completed assessment report. This sign-off includes agreement on the activities required for data quality improvement as noted in the action plan in the assessment report. Once management has approved the data quality assessment report, a copy is to be sent to the Data Quality department and then is posted to the Data Quality intranet page.

**NEW**

In addition to this regular review and support by Data Quality staff, there is a challenge function performed each month on a selected data holding and its most recent data quality assessment report. A few members of the Data Quality staff, as well as two external reviewers (normally managers of other program areas), get together with the manager and team of the program area that is having its most recent report reviewed. Informal discussions occur between those attending the challenge function on topics ranging from the proper application of the assessment tool and sharing good data quality practices to how the Data Quality Framework can be improved. To foster communication of data quality practices and tools across the organization, the selected data holding presents highlights of its latest application of the assessment tool to an internal forum, following the challenge function meeting.

### 4.1.4 Suggested Content and Format

The assessment report must contain enough information in the response to each criterion to justify the rating. This allows those who are unfamiliar with the data holding to understand the rationale for the rating. Similarly, every criterion rated as *not applicable* requires an explanation of why it is not applicable.

The report should begin with a brief introduction that highlights the purpose of the data holding, its use, a description and the time frame for the assessment. Each criterion should be assessed in the order that is laid out in the assessment tool. Recommendations are a key component of the evaluation process. These should be included in relation to problematic criteria or characteristics within the body of the report. They should also be summarized in the executive summary. A Word version of a blank assessment report template is available in Appendix D.



## 4.2 Data Quality Documentation for Users

### 4.2.1 Purpose

The purpose of data quality documentation for data users is to give them sufficient information to decide if the quality of the presented information is fit for their 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. Such data quality documentation does not have to appear in an annual report, although it could be ideally situated there depending on the frequency and types of results presented. For some data holdings that produce an annual report, having the data quality documentation in a methodology notes section or a data limitations section of the report will meet the needs of the users. However, it is recommended that for some data holdings a complete stand-alone data quality document with all the pertinent information that a user needs be created. Any data holding 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.

### 4.2.2 Suggested Content and Format

The data holding's data quality documentation for users should contain enough information to provide the users of the data with a thorough understanding of all limitations associated with the data. The Data Quality department is available to provide input into the development of the report and to review the final version.

What follows is suggested content. Each data holding may develop its own format for its data limitations document; however, this report should contain the following headings or topics:

#### i) Introduction

The introduction should provide the purpose of the CIHI data quality document, including one to two paragraphs describing the data that is 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.

#### ii) Concepts and Definitions

##### a) Mandate or Purpose

The mandate or purpose of the data holding should be given.

##### b) Population

The population of the data holding should be defined, typically clarifying the population of reference and adding a specific time period.

##### c) 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.

### **iii) 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 data holding that was identified as having 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.

### **iv) 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.

### **v) 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. Rather, this section focuses on explaining how the applied collection procedures, etc. 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.

### **vi) Major Methodological Changes From Previous Years**

This section highlights where the current data holding 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 data holding 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.

### **vii) Revision History**

This section highlights changes or corrections made to data that has 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 provided.

### **viii) 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.

### 4.3 Metadata Documentation

A third type of documentation that is critical for any data holding is its metadata. For efficient documentation of methodological processes, data collection, data processing and data dissemination activities, as well as data limitations and data quality measures with respect to a data holding, should be reported in *one well-maintained and exhaustive* source. CIHI has developed a metadata repository that houses a data holding's data quality documentation as well as other valuable metadata about the data holdings.

**NEW**

The metadata repository should be the reference that is first used for questions about specific details on a data holding. One purpose of the metadata documentation is to describe the data flow of a data holding to the point that a person new to the project could take over any particular activity. Information should be included on the steps for data processing, who is responsible, who should sign off on the steps and what data quality checks are in place. The documentation about policies and procedures used within a data holding promotes activities for process improvement, audit trails and specification of responsibilities, and ensures increased reliability in the data. Good documentation allows the data holding to become more process dependent than person dependent, which can result in an increase in data quality and stability.

The documentation for the metadata repository should encompass seven categories that are defined as follows.

1. *Data holding description*—contains essential background information on a data holding to identify sources of the data, uses and users of the data, as well as the responsibilities of those involved in submission, processing, analysis and dissemination.
2. *Methodology*—includes details on the population of interest, population of reference, sample selection (in the case of a survey) and the frame, as well as information on implementing and recording frame changes.
3. *Data collection and capture*—contains procedures and/or specifications for the data being collected or submitted and formatted or standardized for processing purposes.
4. *Data processing*—includes data processing procedures from the time the data is captured by or at CIHI until the analytical file is produced, such as all editing, derivation and estimation processes.
5. *Data analysis and dissemination*—outlines all procedures for analysis as well as the steps in producing or disseminating tables, reports and publications, and responses to queries.
6. *Data storage*—relates to the storage of the data, results and/or estimates from the data holding.
7. *Documentation*—consists of all relevant documentation detailing the data quality achieved and on improving or maintaining data quality.

It should be noted that all seven of these categories are divided into a number of subcategories (118). Detailed information on the subcategories that are included in each of these metadata categories is located in Appendix E.



## Appendix A—Significant Changes From 2005 Version

**NEW**

Below are the significant changes incorporated into this latest version of the Data Quality Framework compared to the 2005 version. A detailed comparison of changes will be produced and published separately.

1. The action plan, which is a deliverable arising from the application of the assessment tool (described in Section 3.5), was expanded to create a useful tool for improving the data quality of data holdings by assigning responsibilities to the recommendations made. A template to create the action plan was also included in this version.
2. The framework was made more applicable for those data holdings that are created from a survey of a sample selected from the population of reference.
3. Updates to certain sections were made to account for those data holdings that have voluntary data submissions.
4. Updates to certain sections were made to account for those data holdings that are longitudinal in nature.
5. The framework was modified to reflect the fact that a new CIHI data dictionary is in existence and that data elements under the object class of provider have been finalized.
6. The development of a standard for metadata documentation is incorporated into this version. An electronic repository for CIHI internal staff (currently called the Master Metadata e-Repository) has been put in place to house all the important metadata documentation.
7. Sections on coverage, capture and collection, unit non-response, measurement error, edit and imputation, and processing and estimation were further expanded to clarify the concepts within these sections. Three criteria within these sections were expanded to have two parts (part a and part b); thus, the total number of criteria increased from 58 to 61.



## Appendix B—Summary of Dimensions, Characteristics and Criteria

The dimensions of data quality and their respective criteria for the 2009 CIHI Data Quality Framework are detailed below. More information on each of the criteria within each data quality dimension and how each criterion is to be assessed for the data quality assessment report is provided in the Data Quality Framework assessment tool (see Appendix C).

<b>Dimensions/ Characteristics</b>	<b>Criteria</b>
<b>Accuracy</b>	
<b>Coverage</b>	1a The population of reference is explicitly stated in all releases.
	1b Efforts are being made to close the gap between the population of reference and the population of interest.
	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>	5a CIHI practices that minimize response burden are documented.
	5b CIHI has documentation of data-provider practices that minimize response burden.
	6 Practices exist that encourage cooperation for data submission.
	7 Practices exist that give support to data providers.
	8 Standard data-submission procedures exist and are followed by data providers.
	9 Data-capture quality control measures exist and are implemented by data providers.
<b>Unit Non-Response</b>	10 The magnitude of unit non-response is mentioned in the data quality documentation.
	11 The number of records for responding units is monitored to detect 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>Dimensions/ Characteristics</b>	<b>Criteria</b>
<b>Accuracy</b>	
<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 and any invalid data is flagged.
	19 Edit rules and imputation are logical and applied consistently.
	20 Edit reports for users are easy to use and understand.
	21 The imputation process is automated and consistent with the edit rules.
<b>Processing and Estimation</b>	22 Documentation for all data processing activities is maintained.
	23 Technical specifications for the data holding are maintained.
	24 Changes to a data holding's underlying structure or processing or estimation programs have been tested.
	25 Raw data, according to the CIHI policy for data retention, is saved in a secure location.
	26a Aggregated statistics from a data holding have been compared, where possible, to similar statistics from another CIHI data holding or external source.
	26b The variance of the estimate, compared to the estimate itself, is at an acceptable level.
<b>Timeliness</b>	
<b>Data Currency at the Time of Release</b>	27 The difference between the actual date of data release and the end of the reference period is reasonably brief.
	28 The official date of data release was announced before the release.
	29 The official date of data release was met.
	30 Data processing activities are regularly reviewed to improve timeliness.
<b>Documentation Currency</b>	31 The recommended data quality documentation was available at the time of data or report release.
	32 Major reports were released on schedule.



<b>Dimensions/ Characteristics</b>	<b>Criteria</b>
<b>Comparability</b>	
<b>Data Dictionary Standards</b>	33 All data elements are evaluated to determine their inclusion within the CIHI Data Dictionary.
	34 Data elements from a data holding that are contained within the CIHI Data Dictionary must conform to dictionary standards.
<b>Standardization</b>	35 Data is collected at the finest level of detail practical.
	36 For any derived data element, the original data element remains accessible.
<b>Linkage</b>	37 Geographical data is collected using the Standard Geographical Classification (SGC).
	38 Data is collected using a consistent time frame, especially between and within jurisdictions.
	39 Identifiers are used to differentiate facilities or organizations uniquely for historical linkage.
	40 Identifiers are used to differentiate persons or machines uniquely for historical linkage.
<b>Equivalency</b>	41 Methodology and limitations for crosswalks and/or conversions are documented.
	42 The magnitude of issues related to crosswalks and conversions falls into one of the predetermined categories.
<b>Historical Comparability</b>	43 Documentation on historical changes to the data holding exists and is easily accessible.
	44 Trend analysis is used to examine changes in core data elements over time.
	45 The magnitude of issues associated with comparing data over time falls into one of the predetermined categories.
<b>Usability</b>	
<b>Accessibility</b>	46 A final data set is made available per planned release.
	47 Standard tables and analyses using standard format and content are produced per planned release or upon request.
	48 Products are defined, catalogued and/or publicized.
<b>Documentation</b>	49 Current data quality documentation for users exists.
	50 Current metadata documentation exists.
	51 A caveat accompanies any preliminary release.
<b>Interpretability</b>	52 A mechanism is in place whereby key users can provide feedback to, and receive notice from, the data holding program area.
	53 Revision guidelines are available and applied per release.

<b>Dimensions/ Characteristics</b>	<b>Criteria</b>
<b>Relevance</b>	
<b>Adaptability</b>	54 Mechanisms are in place to keep stakeholders informed of developments in the field.
	55 The data holding is developed so that future system modifications can be made easily.
<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 assessed.

# Appendix C—Data Quality Framework Assessment Tool

This appendix contains the CIHI Data Quality Framework 2009 assessment tool, which provides descriptions for each of the 5 dimensions, 19 characteristics and 61 criteria that are contained in the CIHI Data Quality Framework 2009. This tool has been created to help data holding staff better understand the framework and to complete the data quality assessment report for their data holding.

This tool is organized into separate sections for each of the five data quality dimensions.

1. Accuracy
2. Timeliness
3. Comparability
4. Usability
5. Relevance

## 1. Accuracy Dimension

Accuracy is what most people think of when they think of data quality. Accuracy refers to how well information in or derived from the data holding reflects the reality it was designed to measure. Thus, when this occurs, the information or data can be considered accurate or reliable.

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

- ✓ Is 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 is the data?

Two characteristics address this issue.

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

✓ What is done with the data?

Two characteristics address this issue.

- Edit and imputation—are the checks and any modifications to the data logical and consistent?
- Processing and estimation—are the processes used to generate values fully tested and documented?

By taking into account the above questions, the accuracy dimension has been divided into seven characteristics. These characteristics are divided into 29 criteria, which are rated individually. It should be noted that three of the criteria have two parts (part a and part b).

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



### Levels of Observation

The calculations in the first three characteristics of the accuracy dimension (coverage, capture and collection, and unit non-response) can be performed at more than one level of observation. There can be levels for frame units and there can be levels within each frame unit, which we call units of analysis. Depending upon the situation, it is possible that a frame unit and a unit of analysis can be one and the same. The following table displays examples of how different numbers of observation levels can result.

Table B Examples of Levels of Observation

Example	Type of Data	Example	Data Provider	Levels of Observation	Comment	Unit	Item
1	Aggregate	Number of graduates	Medical colleges	Frame units (medical colleges), units of analysis (medical colleges)	Equal since aggregate data is the summation of individual records. Aggregated by medical colleges.	Medical colleges	Aggregate data file, for example, number of graduates
2	Individual record	Nurse registration	Nursing associations	Frame units (nursing associations), units of analysis (individual nurses' records)	Frame unit level separate from unit of analysis level. Need separate assessments for criteria 1 to 12.	Nursing associations, nurse's registration	Data on nurse's registration, for example, year of graduation
3	Individual record	Inpatient charts	Ministries of health or hospitals	Frame units (ministries of health or hospitals), units of analysis (patient records)	Two levels of frame units and one level of unit of analysis. Need separate assessments for criteria 1 to 12.	Ministries of health, hospitals, patient records	Data on patient record, for example, gender, admission date

**NEW**

In situations where there is more than one level of observation and an accurate assessment of the data and information quality on all levels is desired, it is essential that all levels be taken into account in the assessment. The overall assessment for a criterion should be based on the *worst* assessment of the applicable levels.

## Coverage

### Criteria

- 1a *The population of reference is explicitly stated in all releases.*
- 1b *Efforts are being made to close the gap between the population of reference and the population of interest.*
- 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.*

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

The **population of interest** is the group of units 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, if information for the complete population of interest is not available, then one must reference it using a population that is available. The **population of reference** is this available group of units and it is the one for which statements are made.

**Example A:** 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 admitting patients on January 1 of the reference year.*

The population of reference for a data holding should be coherent and consistent and, ideally, as close to the population of interest as possible.

The **frame** for a data holding is a list of units (provinces, institutions, doctors, etc.) that provides access to units in the population of reference that will be part of data collection. As mentioned in Table B, it is possible for a data holding to have more than one level of units (that is, frame units, units of analysis). Both administrative databases and surveys 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.**

**Under-coverage** occurs when part of the population of reference is not included among the units on the frame. From Example A, under-coverage would occur if some units of analysis in British Columbia were not included on the frame when they should have been. Under-coverage would occur if a health region in Nova Scotia split into two health regions, but only one was included on the frame. It should be noted that under-coverage can cause a loss of information (that is, lower totals) as well as causing the results to be biased if the units not listed on the frame differ from those that are listed.

**Over-coverage** occurs when units that are not part of the population of reference are included on the frame or when duplicates appear in the database.

Using Example A, data received from the following would result in over-coverage, unless measures were taken to correct the data:

1. A hospital in Saskatchewan
2. An Ontario hospital with no acute care beds
3. A Quebec hospital that opened on March 1

Over-coverage would also occur if data for the same units was submitted by both the hospital and the province, resulting in duplication. Table C provides a summary of the terms defined in this section on coverage.

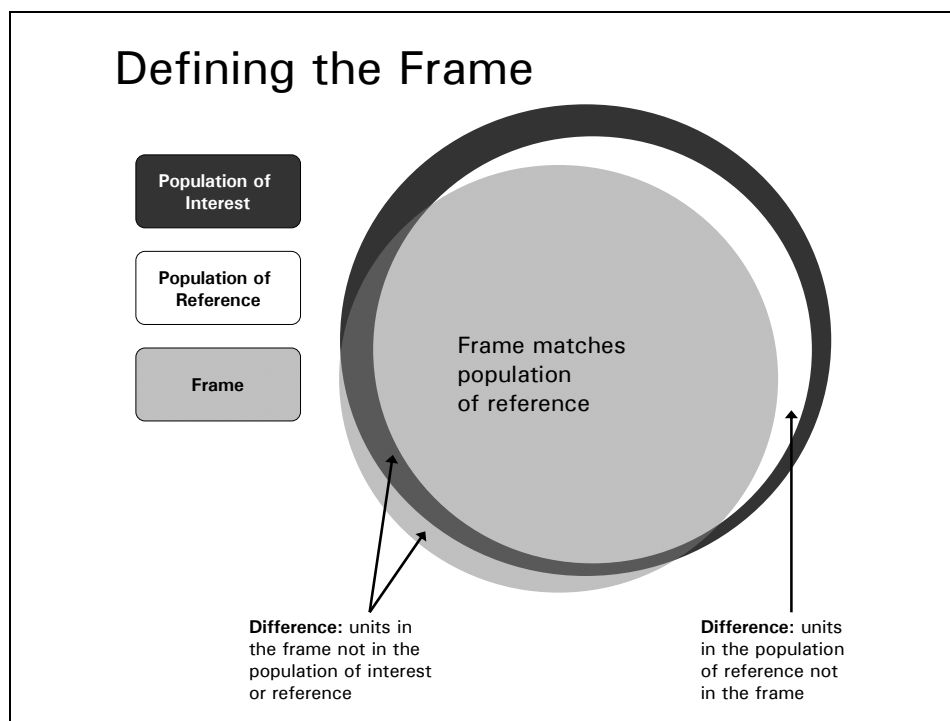
**NEW**

**Table C Summary of Coverage Terms**

Term	Definition
Population of Interest	Data CIHI would ultimately like to collect
Population of Reference	Data CIHI can realistically expect to receive
Frame	Listing of units in the population of reference from which data will be collected
Over-Coverage	When there is more data on the frame than expected (frame > population of reference)
Under-Coverage	When there is less data on the frame than expected (frame < population of reference)

To further clarify the differences between population of interest, population of reference and frame, Figure 1 provides a visual example.

Figure 1 Defining the Frame



Units that should not be included in the data holding and are not included in the **frame** are considered **out of scope**.

Because the use of a good 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 those practices that are used to update the frame.

It is important to realize that coverage errors, with the exception of duplicates, do not necessarily relate to the submission of data. If a unit 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 providers, 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 providers are either added to or removed from the frame).

The degree of under- and over-coverage in a population of reference is determined through the calculation of the rate of under-coverage and the rate of over-coverage.



For example, if data were collected for a defined set of units with the aid of a frame or frames, the **rate of under-coverage** (expressed as a percentage) would be

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

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

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

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

#### Example: Calculating Coverage Error

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

The population of reference is determined by taking into account under-coverage and over-coverage occurrences in the available frame. Thus,

$$\begin{aligned} \text{Population of reference} &= (\text{no. on the frame}) + (\text{no. missed}) - (\text{no. erroneously included}) \\ &= 1,100 + 25 - 5 = 1,120 \end{aligned}$$

$$\text{Under-coverage rate} = \frac{\text{no. missed}}{\text{population of reference}} \times 100 = \frac{25}{1,120} \times 100 = 2.2\%$$

$$\text{Over-coverage rate} = \frac{\text{no. erroneously included}}{\text{population of reference}} \times 100 = \frac{5}{1,120} \times 100 = 0.4\%$$

#### Coverage When There Are Voluntary Data Submissions

**NEW**

The preceding paragraphs deal with coverage when submission is mandated. When submission is voluntary, the difference between the **population of interest** and the **population of reference** can often be large. One should not try to draw conclusions on the population of interest when data is missing. Make statements only about the population of reference, that is, those that are voluntarily submitting. For example, if the population of interest consisted of all medical equipment in one province then, where only data about the oldest pieces of medical equipment is voluntarily submitted, one cannot make statements about the entire population of interest, since no data was received about the newest pieces of medical equipment.

In such a situation involving voluntary data submissions, the population for which data is expected needs to be clearly defined. Once the frame is established for those that have agreed to provide data and a profile containing essential information is created for them, it is possible to determine the population of reference and, thus, compute over- and under-coverage rates. It is important to carry out regular frame maintenance procedures to

determine if the units are still in scope for the population of reference. In addition, frame changes over time involving additions and deletions should be noted in documents such as the data quality assessment report (see Section 4.1) or the data quality documentation for users (see Section 4.2).

**Criterion 1a** *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. **Any difference between the population of reference and the population of interest should be made known to the users in the release.** A release is any report, data release or output from the data holding. The release should explicitly state whether the population of reference has been created from mandated or voluntary submissions. In the case of voluntary submissions, any important agreements should also be explicitly noted. Any minor issues or clarifications are to be mentioned in the footnotes, at the very least.

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

This criterion is met only if the population of reference is stated in *all* releases that have taken place since the last data quality assessment or since data releases have been occurring.

**NEW**

**Criterion 1b** *Efforts are being made to close the gap between the population of reference and the population of interest.*

In cases where any difference between the population of reference and the population of interest has been discussed, mention should also be made of any efforts undertaken by the program area to close the gap between these two populations.

This criterion is met if the program area has activities in progress or has completed efforts to close any existing gap to the extent possible.

**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. When documenting under- and over-coverage, each level of observation (that is, frame unit) should be mentioned as described in Table B. If the under- or over-coverage can be corrected, the data should be corrected before results are published and the source of the coverage error documented for internal use. Over-coverage can be corrected by removing the out-of-scope or duplicate units from the frame, while under-coverage can be corrected through the use of an independent source that will identify missing or new units. If the data cannot be corrected, the known 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 and 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 data holding. 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 to ensure that the frame is up to date (for example, to confirm that the correct continuing care facilities are on the frame, a comparison should be done with each of the provincial ministries of health). 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 and colleges of medicine, nursing or pharmacy. If an independent data source cannot be found, the frame should, at a minimum, be verified by such sources.

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 were facilities with MRI machines in New Brunswick, and we could only obtain 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 and the problem will need to be resolved before proceeding further.

This criterion is met if the frame has been compared to external sources on an annual basis to determine the presence of errors on the frame.

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

This criterion is designed to put some qualitative measure or rating on the effect of under- or over-coverage. Consider the following guidelines:

Possible Rating	Rate of Under- or Over-Coverage (%)
None or minimal	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 rates separately and use whatever rating is the lower of the two. For example, a database with minimal over-coverage and significant under-coverage should be rated as *significant* for this criterion. Consider as well the different levels of observation when calculating the under- and over-coverage. Each frame unit should be assessed individually as described in Table B.

**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 still has a coverage error. Out-of-scope or duplicate units do not compensate for units that are missing on the frame.

## Capture and Collection

### Criteria

- 5a) *CIHI practices that minimize response burden are documented.*
- 5b) *CIHI has documentation of data-provider practices that minimize response burden.*
- 6 *Practices exist that encourage cooperation for data submission.*
- 7 *Practices exist that give support to data providers.*
- 8 *Standard data submission procedures exist and are followed by data providers.*
- 9 *Data-capture quality control measures exist and are implemented by data providers.*

The capture and collection characteristic refers to the practices that are used when dealing with the data providers and during data entry. The **data provider** is the person or organization that provides the data to the data holding. The data providers 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 by CIHI from different data providers into a common data holding.

As an example for a mental health data holding, the data provider may be defined as the mental health facility and/or ministry of health: data capture is the abstraction of hospital charts by the mental health facility or the transfer of such charts from the facility to the ministry of health, and collection is the receipt of the electronic file by CIHI directly from the facility or from the ministry of health.

Current technology allows the capture of information to be done in an automated fashion using software developed internally or by an outside vendor. The qualities of an optimal capture software program are the following:

- Only information on necessary data elements is collected (for example, no need to collect age if date of birth is collected);
- Mandatory data elements are forced to be captured;
- Default values are not set. It is necessary to distinguish a value that is not present but should be because it is missing or unknown (non-response) from one that is not applicable;
- Drop-down menus exist to minimize any possible keying errors;
- Built-in edits exist with meaningful error messages to locate errors and make modifications rapidly;
- It has flexibility to adapt to new requirements (if needed); and
- In the case of an outside vendor, all definitions and valid values are provided through the vendor's software manual.

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 regard to data providers, the question that should be asked is "What can we do to make the data provider's job as easy as possible, while still getting the information that we need?"

**Criterion 5a** *CIHI practices that minimize response burden are documented.*

This criterion assesses whether measures are used to ensure that the effort required by the data provider is minimized without compromising data quality. This is otherwise known as minimizing the **response burden**. There are many ways this can be done: electronic capture and submission (for example, using capture software), reasonable submission schedules, exclusion of unnecessary data elements, etc. However, it should be noted that there may be situations in which the data provider collects data for its own purposes; thus, the program area may not get all the information it desires without additional requests to the provider.

This criterion is met if documentation of the implemented CIHI practices to minimize response burden exists.

**NEW**

**Criterion 5b** *CIHI has documentation about data-provider practices that minimize response burden.*

This criterion assesses whether measures are used by an organization (those where CIHI is not directly capturing or collecting the data from the initial source) to capture the data, ensuring that the initial source of the data has to do as little unnecessary work as possible. This is otherwise also known as minimizing the **response burden**. There are many ways this can be done: electronic capture and submission (for example, using capture software), reasonable submission schedules, exclusion of unnecessary data elements, etc.

This criterion is met, for those instances where CIHI is not directly capturing or collecting the data from the initial source, if such documentation is accessible by CIHI. If CIHI does not have access to such documentation or it is not available from the data provider(s), then the criterion is not met.

**Criterion 6** *Practices exist that encourage cooperation for data submission.*

Practices that encourage cooperation are important whether the submission of data is voluntary or mandated. However, when data providers have the option of not responding, a collaborative effort will be necessary to receive quality data.

Practices to encourage the data provider to submit data can be as simple as stressing the importance of participation, the assurance of confidentiality, the provision of compensation for cooperation (for example, free publications, training and specialized reports) or the acknowledgement of the receipt of data in a letter or email thanking the supplier for the information. The latter two will aid in encouraging future cooperation.

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

**Criterion 7** *Practices exist that give support to data providers.*

Providing support to data providers is essential to ensure that data is submitted promptly and correctly. This support can be done before data capture and during data capture. Prior to data capture, education sessions and training can be administered, as well as having persons in place to promptly respond to emails and phone calls from the data providers. During data capture, emails and phone questions should still be answered promptly. In addition, technical and coding support should be made available to data providers by providing access to supporting documentation, coding guidelines and the abstracting manual.

This criterion is met if there are methods both before and during data capture by which support is given to the data providers.

**Criterion 8** *Standard data submission procedures exist and are followed by data providers.*

Standard data submission procedures make the data collection process easier for both the data provider and internal personnel at CIHI. They may also result in improved timeliness and reduced errors.

Standard submission procedures ensure that data collection is done as consistently as possible across suppliers. A data holding in which some suppliers submit annual data on paper and others submit monthly data electronically does not have standard procedures. In the same regard, data elements that are collected across jurisdictions should be the same as well as having similarities among mandatory and optional data elements. This will facilitate meaningful comparisons across jurisdictions.

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

**Criterion 9** *Data-capture quality control measures exist and are implemented by data providers.*

**Data-capture quality control** measures are carried out when those responsible for data capture are entering the data. These measures ensure that the data is 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 is present. Data-capture edit checks generally consist of validity edit checks by data element (for example, checking to see if the gender is reported as *male*, *female*, *other* or *unknown*, or if the patient identification number is the appropriate length). In addition, consistency edit checks verify the relationship between data elements (for example, verifying, through a capture system, an intervention that can only be performed on females). Edit checks must also be applied when the data is originally loaded (refer to criteria 18 to 21).

**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. It should be noted that the Data Quality department is available to assist program areas in setting up the parameters for such verifications.

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

## Unit Non-Response

### Criteria

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

**Unit non-response** occurs when responses for entire units (province, health region, facility or patient records, etc.) are missing from the data holding. Unit non-response is often confused with under-coverage, as both occur when complete units are missing. For instance, if a unit on the frame for a data holding does not submit data, it is a case of a frame unit being a non-response, whereas under-coverage would occur if the unit was not included on the frame but is in the population of reference. When examining frame units, review the different levels of observation in Table B to determine which units you will report here.

The nature of administrative databases often results in a data-collection process through data providers where non-response can occur at various levels (that is, patient, service provider, facility, health region or province). For instance, for drug claims data, if ministries of health are expected to submit data for all drug claims in their public drug programs (population of reference), data from some programs may not be submitted at all, while only a portion of the drug claims may be submitted under a particular program. These are both examples of unit non-response. It is important to get some measure of the amount of data not submitted by a data provider. Unit non-response should be kept to a minimum in order to

- Minimize the loss of valuable information and
- Avoid biased results when units for which no information was received differ from those from which information was received.



At CIHI, a **unit non-response rate** is often calculated rather than its complement, the unit response rate. The response rate calculation can be based on the number of **frame units** or the number of **units of analysis**. The frame unit is the unit that is on the frame for the data holding (for example, patient, service provider, facility, health region, province or other data provider), whereas the unit of analysis can also be the frame unit or a unit that is within a frame unit (for example, patients within a facility). See Table B for further examples of these levels of observation. The **frame unit** non-response rate (expressed as a percentage) is calculated as follows:

$$\frac{\text{no. of frame units that did not submit data}}{\text{no. of units on the frame}} \times 100$$

The non-response rate for the **units of analysis** over all the frame units (expressed as a percentage) is

$$\frac{\text{no. of units that did not submit data for each of the frame units}}{\text{no. of units that should have submitted data for each of the frame units}} \times 100$$

The unit response rate (expressed as a percentage) is simply equal to 100% – unit non-response rate.

Depending on the data holding, it may be appropriate to calculate a non-response rate on sub-groups, such as size, type, region or province, in addition to calculating an overall non-response rate. Missing data from a large frame unit can result in the loss of many more units of analysis than missing data from a small frame unit. The approximate non-response rate for the units can be calculated by comparing the overall number of units of analysis not received to the number of units of analysis expected. This calculation can become quite complex. Consultation with Data Quality department staff 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 frame units have provided all the required units of analysis. It is, however, relatively easy to detect large changes in the number of units that a frame unit submits. Significant changes can give an indication of a high non-response rate or the inclusion of incorrect units and should be examined. The number of units of analysis not received from each frame unit should be monitored on an ongoing basis so that unusual numbers can be examined.

Through comparisons with previous years or from the monitoring done, any unit non-response that is detected can best be resolved through follow-up actions. For example, a frame unit should be followed up when nothing is submitted by it to make sure that the frame unit is still in scope. Also, follow-up should be performed for responding frame units when unusual numbers for units of analysis are detected by comparison with a previous period.

Unit non-response rates can be minimized through good communication (education sessions, training, abstract manual, etc.) with data providers before the collection period starts. In addition, the response burden on data providers can be reduced by inquiring how their job can be made easier and acting upon it. It is also important to learn from any mistakes encountered and have them documented. In addition, giving the providers of data sufficient time both to submit data and to begin follow-up activities with them will aid in minimizing unit non-response.

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

It is important that the magnitude of the frame unit non-response be reported in the data quality documentation so that users can assess the completeness of the data. The frame unit could be the patient, service provider, health region, province or other data provider as described above. If non-response rates vary significantly by province or region, the non-response rate should be reported at these levels.

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

**Criterion 11** *The number of records for responding units is monitored to detect unusual values.*

In order to detect non-response, it is important that the number of units responding be tracked over time to detect unusual values. For example, if a data provider submits records monthly for approximately 1,000 units, a detailed investigation should occur if the number of records received suddenly jumps to 1,500 or drops to 500 units. The change in the number of units submitted does not necessarily indicate problems, as there are many possible reasons for the change (late submission of records, expansion of the institution or temporary closure, new hiring practices, an epidemic causing reductions in hospital visits, etc.).

This criterion is met if the numbers of records for responding frame units 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 both the frame level and the unit of analysis level. It is suggested that the non-response be calculated for each level of observation, frame and/or unit of analysis. As indicated previously, a unit non-response rate is usually computed at CIHI rather than its complement, the unit response rate.

The unit non-response rate at the frame level (expressed as a percentage) is

$$\frac{\text{no. of frame units that did not submit data}}{\text{no. of units on the frame}} \times 100$$

The unit non-response rate at the unit of analysis level (expressed as a percentage) is

$$\frac{\text{no. of units of analysis that did not submit data}}{\text{total no. of expected or known units of analysis}} \times 100$$

The following table contains suggested ratings for the unit non-response rate for the frame unit or unit of analysis level. If there is substantial non-response above a certain desired percentage because the frame units that did not report are very large, the ratings can be adjusted to reflect more accurately the severity of the problems created by the missing data. Ideally, the unit non-response rate should be determined by those who best know the units on the frame (either the data provider or CIHI staff). For those situations where CIHI is not knowledgeable of all the units in the frame, it is important that CIHI gather the relevant information (for example, respondent status) on each unit from the data provider in order to correctly determine the unit non-response rate.

Suggested Rating	Unit Non-Response Rate (%)
None or minimal	Less than 2%
Moderate	2% to 10%
Significant	More than 10%

### Item (or 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 unit (either frame unit or unit of analysis) for which data has been received contains only partial information for the data elements or transactions (in financial data holdings). Item non-response does not include those data elements that are not applicable for the unit, in other words where information for some of the data elements is missing. Missing refers to such data elements where information was expected and instead is blank on the data file. A data item is considered missing even if default values (that is, 0 or 99) are used to code a non-reported item as a non-response. For item non-response purposes, data elements with a reported zero value should not be considered to be missing. Item non-response differs from unit non-response in that unit non-response deals with response from entire units (as described earlier in criteria 10 to 12), while item non-response deals with the specific data elements within the units.

Item non-response rates are quite simple to calculate when it is possible to identify when the data elements are missing as opposed to being not applicable. If hospitals were asked the number of psychiatric beds they have, a blank answer may mean that the institution has no psychiatric beds (not applicable) or that it simply did not respond (missing).

During the editing process it is recommended that an additional data element be used to distinguish those elements that are missing values from those that are not applicable. For instance, the standard accepted within the CIHI Data Dictionary is that missing is coded as either 7 (not collected) or 9 (unknown) while 8 is the code for non-applicable. If an additional data element is not in place, it is important that non-applicable responses and missing responses can be differentiated prior to calculating the item non-response rate, as non-applicable cases are to be excluded from the calculation.

At CIHI, an **item non-response rate** is often calculated rather than its complement, the item response rate.

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

$$\frac{\text{no. of units for which data for a data element was not provided}}{\text{no. of units that should have provided the data element}} \times 100$$

It should be noted that units (either frame units or units of analysis) containing a data element that is not applicable for that unit are to be excluded from both the numerator and denominator in that particular data element's calculation of item non-response rate.

The following is an example to further illustrate the concept of item non-response. One hundred units (that is, facilities) from a data holding's frame were asked two questions: How many pediatricians work at your facility? How many nurses work at your facility? The response to these questions yielded that 70 of these facilities have both pediatricians and nurses and 5 have only nurses (never have pediatricians). The other facilities—which are known to have both pediatricians and nurses—failed to submit anything at all.

Thus, the item non-response rate for the number of pediatricians is  $25/95 = 26.3\%$  (5 units without pediatricians excluded from both numerator and denominator), while the item non-response rate for the number of nurses would be  $25/100 = 25\%$ .

It is important to realize that the item non-response rate is calculated in comparison to the number of units that report the data element (where the data element of the unit is applicable). In this example, although we have a 26.3% item non-response rate for the number of pediatricians, we do not have pediatrician data from 30% (30/100) of the facilities. The item non-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 0% item non-response rate for all data elements may be missing much data if the unit non-response rate is high. Similarly, a database with a low unit non-response rate and high item non-response rate will also be missing much data, especially given that complete units are rendered useless for analysis purposes when they have many missing data elements.

This criterion is considered met if item non-response rates are calculated for all core data elements (those that are routinely used in analyses 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 as mentioned above to be able to distinguish between blank values and non-response.

Data elements are normally flagged during the editing process by the creation of an additional value for the existing data element (for example, any value that is not a possible response value), or the creation of a new data element altogether. For instance, the standard accepted within the CIHI Data Dictionary is that missing is coded as either 7 (not collected) or 9 (unknown) while 8 is the code for not applicable. This allows easy identification of a missing or non-applicable data element.

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 analyses of the data.

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

This criterion focuses only on the volume or magnitude of data elements with item non-response. The effect of a certain magnitude of item non-response depends on many factors, including the importance of the missing responses to data elements and/or core data elements, and whether there is any pattern to the missing values. Ideally, the missing values should be “missing completely at random” to avoid bias in the results. Data is considered missing completely 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. When rating the level of item non-response, the core data element with the highest item non-response rate should be considered. The following ratings are only suggested ratings. If certain data elements are especially important or it is determined that they are not missing completely at random, the rating may be changed to reflect more accurately the severity of the problems created by the data that is missing for the data elements.

The suggested ratings are the following:

Suggested Rating	Item Non-Response Rate (%)
None or minimal	Less than 2%
Moderate	2% to 5%
Significant	More than 5%

## Measurement Error, Bias and Consistency

### Criteria

- 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.*

Errors can occur for many different reasons. It can often be difficult to group the errors that do occur to allow for an easy assessment of the differing causes. As a way to simplify this, the framework has divided the assessment of errors into three overlapping components:

1. *Measurement error*—error caused when a data element is coded or answered incorrectly.
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.

These three components are further explained in the following paragraphs.

High **measurement error** can indicate a number of issues to be resolved. These can include having unclear definitions, lack of training causing numerous interpretations and variability of responses for subjective entries, over-editing of the data or a weakness in the data-collection procedure. The implementation of automated procedures that are fully tested and reviewed, along with good documentation and training, will help in reducing any measurement error.

The *existence* of **bias** can be very hard to prove without special studies such as a reabstraction study (described later), although it can be relatively easy to detect *possible* biases. Possible biases can be detected when the instances of not having complete coverage or complete response or when processing or sampling errors are known to be occurring. Coverage errors, which consist of omissions, erroneous inclusions and duplications in the frame, can cause either a positive or negative bias in the data, and the effect can vary for different sub-groups of the population of interest. Non-response errors occur when the data holding fails to get a response to one, or possibly all, of the questions and can cause a bias if non-respondents and respondents differ with respect to the characteristic of interest. Processing error, which can occur at subsequent steps following capture and collection, such as in data editing and tabulation, can lead to biased results being produced. In the case of a random sample being selected from the population of interest, bias can be introduced to the estimates, due to the fact that the survey is based on a sample of the population rather than the entire population, or if a proper probability sample is not chosen. In addition, with a sample survey, estimation errors may introduce bias depending upon the choice of the estimation method being used.

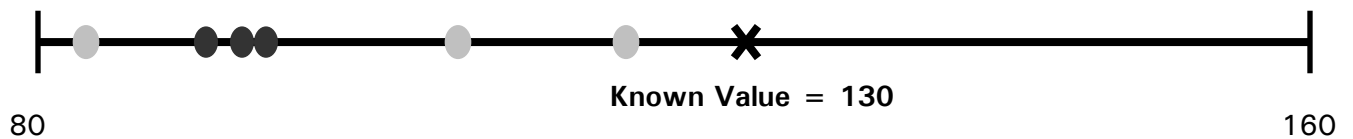
When considering bias, it is important to consider **correlated bias**, which is a bias that results when one data element is correlated with another data element. For example, correlated bias can be introduced when the length of time that individuals are observed is correlated with their outcomes. In such a situation, it is possible that a minimum eligibility period is imposed that leads to the omission of patients with short-term eligibility. If the reasons for the failure to meet the minimum eligibility criteria are correlated with outcomes, such as death, then a biased conclusion can result. Similarly, if the length of the observation period varies for individuals, then bias may be introduced because of failure to observe an outcome after the observation period ended. 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 and differences can be detected.

**Consistency** measures the variation of the responses over repeated measurements and, in some cases, is referred to as reliability. Subjective data elements (such as level of impairment on a scale of 1 to 5 or diagnosis type) are those that may not have a correct answer. Consistency not only applies to subjective data elements, but can also be a factor for data elements where there is an occurrence of measurement error (for example, measuring height incorrectly). The consistency component 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. The kappa statistic (or coefficient) measures the pairwise agreement among a set of coders that are making category judgments, and it corrects for expected chance agreement. The formula is the following:

$$k = \frac{(\text{observed count of agreement} - \text{expected count of agreement})}{(\text{total number of respondent pairs} - \text{expected count of agreement})}$$

The resulting estimates and tests of agreement among multiple coders are appropriate when responses are on a nominal or ordinal scale. Consult Data Quality department staff for information on which technique is most appropriate for the data.

The following example further illustrates the difference between bias and consistency. It shows the results of having two nurses who have each taken a patient's blood pressure three times during a specific time period.



As the illustration indicates, the readings from the nurse indicated by the light grey circles are **less consistent** than the readings from the nurse indicated by the black circles, since the measurements are more dispersed. Thus, the nurse indicated by the black circles is **more consistent** than the one who is indicated by the light grey circles. However, the one indicated by the light grey circles is **less biased** since, on average, the readings for this nurse are closer to the known value than those for the other nurse.

While the exact level of measurement error, bias and consistency of data elements is often not known, special studies such as reabstraction studies should be considered. In a reabstraction study, a sample of records (for example, patient charts) is selected and then recoded. Data that was originally coded is compared to data collected in the study. Any discrepancies between the values and reasons for discrepancies are identified. In a reabstraction study, measurement error is evaluated by the discrepancy rate between the database and the reabstractor. Bias, meanwhile, is evaluated by looking at the discrepancy rates as a whole; if the errors are systematic, then the results will be misleading and may require adjustment. Consistency is evaluated by repeated measurements such as inter-rater or intra-rater checks. In an inter-rater reliability check, two reabstractors must code the same chart to determine if the responses are consistent between reabstractors. In an intra-rater reliability check, one reabstractor must code the same chart twice to determine that particular person's consistency in recording information.

A special study such as a reabstraction study is something that cannot always be performed for data releases due to limited resources. However, the personnel working with the data holding often have an idea of the quality of the data elements in the database and can provide an initial assessment of the three types of errors.

**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 reabstraction 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 data elements.

If a reabstraction-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 not important, the rating can be adjusted accordingly.

<b>Suggested Rating</b>	<b>Error Rate for Non-Subjective Variables (%)</b>
None or minimal	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, it does not necessarily mean the criterion is automatically rated as *unknown*. In many cases, data holding personnel may be aware of problems or have an idea of the amount of error in the data without having to conduct a special study. 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 data elements 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 data elements. 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 data element) 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 may exaggerate the amount of pain they are in, which is a subjective data element.

This criterion is difficult to evaluate, due to the complexity of proving whether bias has occurred. The assessment, therefore, is based on whether there is, or is perceived to be, 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 reabstraction. If there is, or is believed to be, 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 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, bias significant enough to affect resulting totals, frequency counts or estimates (in a sampled population scenario).

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

Consistency is a concern for all data holdings, as consistency for each of its data elements is highly dependent on the opinion or interpretation of the coders, nurses and registrants, for example. Consider both the consistency of the measurements from these individuals and the consistency of measurements between them. Similar to bias and measurement error, consistency is most often assessed through reabstraction 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. Consult Data Quality department staff for information on the measurement technique that is most appropriate for the data in question and for further analysis and interpretation.

If a special data quality study has been done, the results for the subjective data elements 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 if the data elements are either especially important or not, the rating can be adjusted accordingly.

Suggested Rating	Discrepancy Rate (%)	Kappa Statistic
None or minimal	0% to less than 5%	0.81 to 1.00
Moderate	5% to 10%	0.50 to 0.80
Significant	Greater than 10%	-1.00 to 0.49

If the level of consistency is not estimated through a data quality study, it does not necessarily mean the criterion is automatically rated as *unknown*. In many cases, staff working on the database may be aware of problems or have a notion of the reliability of the data. 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 are unable to evaluate it qualitatively, the criterion should be rated as *unknown*.

**Edit and Imputation**

**Criteria**

- 18 *Validity checks are done for each data element and any invalid data is flagged.*
- 19 *Edit rules and imputation are logical and applied consistently.*
- 20 *Edit reports for users are easy to use and understand.*
- 21 *The imputation process is automated and consistent with the edit rules.*

At CIHI, **editing** is the application of checks to identify whether units or data elements are missing, invalid or inconsistent. Such identification will point to data items that are potentially in error. **Imputation** is the process of determining and assigning replacement values to resolve problems with data identified at the editing stage as being missing, invalid or inconsistent. Editing and imputation can be built into one process or two separate processes.

Editing of the data can be a complex task. Setting up proper edits is an investment in data quality. A good editing program can identify many errors in the data that might not be detected otherwise. When setting up edits, it is important to consider what type of edit will be appropriate and what the parameters of the edit will be. Following the collection of data, one of the first steps in an editing program should be to identify if there is duplication. Examples of duplication include a facility sending the information twice on the same day or when updated information is submitted for a unit but the previous information is not removed from the database. What are some ways of identifying duplicate units? First of all, specific data elements need to be determined that uniquely identify a unit. Second, it also needs to be verified that the unit is not already in the database when doing any updates. In addition, a log of edit failures and follow-up units that have not been resubmitted should be kept.

Once it has been determined that all duplicates have been resolved, further editing can continue. There are essentially three types of edits to consider: validity edits, consistency edits and distribution edits. Validity edits identify cases where missing values occur for mandatory data elements, where the format for responses is incorrect or where response values do not fall within a specified range. Consistency edits check for the improbable relationships between data elements (for example, a male giving birth to a child). Distribution edits identify outliers with respect to the distribution of the data by looking at the entire data set (for example, wait times being considerably longer than the average). Validity and simple consistency edits can be performed during data capture (incorporated into the abstraction software) while more complicated consistency edits and distribution edits can be performed during data processing. Consultation with Data Quality department staff is recommended if complicated edits are being planned, as well as to verify any parameters being used.

When a unit or data element is flagged by edit checks during data processing, a follow-up with the data provider should be done before the end of the collection period; otherwise, imputation should be considered. After the collection period end, imputation can be performed to determine replacement values where necessary for those values that are missing, invalid or inconsistent.

Following any imputation, it is important to make sure that non-imputed and imputed data can be differentiated. One way to do this is through the creation of an additional data element (either a flag indicating that imputation has been performed or one that contains the imputed values). Furthermore, it is a good habit to monitor the number of records being imputed as well as the number of times an imputation method (historical imputation, mean

imputation, etc.) is used. This may help to identify if modifications need to be made to improve data quality. Any time imputation is performed, it is also recommended that the various edit checks be re-applied to ensure that the imputed unit or imputed data element is error free and internally consistent.

Two common methods of imputation are logical (or deductive) imputation and historical imputation. For logical imputation, only one possible value exists for a missing or invalid situation (for example, if the date of registration and the date visit completed are the same date, the only possible value for the triage date is also the same date). For historical imputation, data provided for a previous period is carried forward for the current period.

**Criterion 18** *Validity checks are done for each data element and any invalid data is flagged.*

Validity checks ensure that the proper response format is used and that 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 done on a data element representing a date, for instance, can either ensure that the response is in an acceptable date format (for example, YYYYMMDD) or 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. For example, if 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 since this date is not valid.

Invalid data in a database will quickly raise questions about the quality of the data, and as invalid data is nearly impossible to justify (how would one explain a patient weighing negative 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 is flagged.

**Criterion 19** *Edit rules and imputation are logical and applied consistently.*

Edit rules are logical if they make sense with regard to the data that is collected. Imputation should not be used to modify data that *may* be correct, but can be used to modify data that is obviously incorrect. It is important to note that a verification of particular edits applied at one stage (either data capture or data processing) does not contradict another edit being applied at a different stage.

This criterion is met if the edit rules and imputation are determined to be logical and if their application is applied consistently to take full advantage of the data being provided.

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

Edit reports should clearly identify the units and/or data elements that passed or failed the edits and the reason for the failure.

In order to have an efficient editing process, it is important that edit reports be easy to understand. In cases where the data is sent back to be modified by the data providers, it is especially important that the reason for the edit failure be clearly reported to the data providers. If the reason is not given, it can be very time-consuming for the data providers to examine these failures manually to determine what modifications are needed.

The number of times a particular edit rule is applied has implications for the quality of the data. If an edit rule is applied 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 any event, the data and the edit rule should be examined and it may result that a resubmission (if possible) from the data provider is required. Finally, any edit report for the users should clearly indicate the time period (for example, date and time stamp) being studied.

This criterion is met if the edit reports are easy to understand and in a usable format.

**Criterion 21** *The imputation process is automated and consistent with the edit rules.*

As defined above, imputation is the process of determining and assigning replacement values where necessary for missing, invalid or inconsistent data. When a non-applicable data element is missing, the use of an indicator flag in that field or another field to indicate that the data element is not applicable should not be considered imputation.

Imputation is an accepted statistical practice when properly done and accounted for. Performing a completely tested and approved method of imputation in an automated fashion is seen as the appropriate practice. Manual imputation is often subjective, difficult to trace, not reproducible and easily questioned (why one value and not another?). Proper automated imputation, on the other hand, is less subjective, easy to trace and easier to justify when done appropriately. In addition, the resulting imputed values should pass the edit rules that were applied to the reported or collected data.

Imputation can have drastic effects on 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 and is consistent with the edit rules.

## Processing and Estimation

Criteria	
22	<i>Documentation for all data processing activities is maintained.</i>
23	<i>Technical specifications for the data holding are maintained.</i>
24	<i>Changes to a data holding's underlying structure or processing or estimation programs have been tested.</i>
25	<i>Raw data, according to the CIHI policy for data retention, is saved in a secure location.</i>
26a	<i>Aggregated statistics from a data holding have been compared, where possible, to similar statistics from another CIHI data holding or external source.</i>
26b	<i>The variance of the estimate, compared to the estimate itself, is at an acceptable level.</i>

**Processing** is the systematic application of programs or procedures to a database for almost any purpose. The application should follow a logical order, with each of the programs being able to stand alone as it performs specific functions. Generally, processing transforms data obtained during collection into a form that is suitable for data analysis or tabulation. For instance, data that is sent to CIHI is often “processed” by going through the following steps:

1. Errors in the units or data elements are identified through the editing process.
2. A decision is made about how units or data elements will be modified—either through follow-up or imputation.
3. Corrections, if any, are made to the data.
4. Data elements are derived or grouped where necessary.
5. Files are created for analytical and/or tabulation purposes.

Keep in mind that any error during any of the steps can have a significant impact on the data quality of a database.

Since errors can occur at any stage of processing, particularly for manual and repetitive activities, processing should be monitored and corrective action taken when necessary in order to maintain or improve quality. This is generally done by implementing quality control and quality assurance procedures.

Quality control procedures are those that are in place to verify the incoming data and the data that is going through the processing systems. If caught early enough in the process, modifications can be quickly performed that will not jeopardize the quality of the data. It should be emphasized that the focus in quality control procedures is the data itself. For example, quality control procedures include checking for duplicate units in the data holding before adding new units and keeping track of edit failures to ensure that data is resubmitted correctly.

Meanwhile, quality assurance procedures focus on the process and include quality control procedures plus a number of other procedures in order to achieve desired quality in the data and the resulting estimates. Examples of other procedures included in quality assurance include the calculation of statistics that assess quality, the proper training of those persons processing the data, the use of test data to verify if new components have been implemented correctly and maintaining good documentation.

**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 from that population. This can involve taking data from either a sample or a census, data that contains less than 100% response or data with coverage issues. Adjustments are then applied to the data, where necessary, such that calculated values represent the population of reference. 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, since they often have some type of error associated with them.

There are various types of errors that can exist in the data. The following identifies five types of errors with the associated way of measuring and reporting the error: four errors are referred to as non-sampling errors and one is a sampling error.

**NEW**

### **Non-Sampling Errors**

1. Coverage error (described in the Coverage section), which is measured through the use of under-coverage and over-coverage rates.
2. Non-response error, which is measured through the use of either non-response or response rates. The sections on Unit Non-Response and Item (Partial) Non-Response provide further information for this type of error.
3. Measurement error (described in the Measurement Error section), which is measured through discrepancy rates.
4. Processing error (explained in the Edit and Imputation section), which is reported via edit failure rates and imputation rates.

### **Sampling Error**

Sampling errors occur when only a portion of the population is selected (or surveyed). If a sample is selected that is representative of the population of reference and the non-sampling error is at a minimum, then the estimate resulting from the sample survey with its associated sampling error should, in the majority of cases, include the true estimate (the one that would have resulted if the responses of all frame units had been obtained).

The three common measures of sampling error are the *variance*, the *standard error* and the *coefficient of variation (CV)*.

1. The variance is the measure of the variability of the estimates obtained when drawing all possible samples from the population of reference. This measure is often not reported when it is a very large number. Standard error and/or CV are the statistics normally reported.
2. Standard error is simply the square root of the variance.
3. CV is the absolute measure of dispersion, and it is measured by taking the standard error and expressing it as a percentage of the estimate as follows:

$$\frac{\text{Standard error} \times 100}{\text{Estimate}}$$

It is highly recommended that Data Quality department staff be consulted when calculating these measures and/or when designing a sample survey.

An example of these three measures of sampling error follows. An estimate was requested of the total number of people diagnosed with diabetes in a particular province by taking a sample survey. It was determined that a representative sample would consist of 2,000 persons, and from responses to this sample it was estimated that there are 10,000 diabetics in the particular province. The variance of this estimate was determined to be 250,000 (normally the measure of the variance is not included in a report since the other measures are sufficient). The standard error and CV can be easily calculated once the variance is determined. For the confidence interval for the estimate, the 95% confidence interval is often chosen; thus, the applicable coefficient to use from a standard normal probability table is 1.96. By referring to such a table, it is possible to determine the applicable coefficient for any desired confidence interval. Once the coefficient is determined, the formula for the confidence interval is simply the following:

$$\begin{aligned} \text{Confidence interval (lower bound)} &= \text{estimate} - \text{coefficient} \times \text{standard error} \\ \text{Confidence interval (upper bound)} &= \text{estimate} + \text{coefficient} \times \text{standard error} \end{aligned}$$

The estimated measures from the above example are summarized in the following table.

<b>Estimate of Total</b>	10,000 diabetics
<b>Variance of the Estimate</b>	250,000
<b>Standard Error</b>	Square root of 250,000 = 500
<b>CV</b>	$(500 / 10,000) \times 100 = 5\%$
<b>95% Confidence Interval</b>	$10,000 \pm 1.96 \times 500 = 10,000 \pm 980$ = [9,020; 10,980]

The 95% confidence interval indicates that there are 95 chances in a hundred that the “true” value for the total number of diabetics in the particular province is between 9,020 and 10,980.



**Criterion 22** *Documentation for all data processing activities is maintained.*

Processing is the sequence of steps used when loading the data, editing the data, producing estimates, etc. Movement of staff outside of the organization or work unit can easily result in a loss of knowledge about the processing steps, which can result in errors in the data. The documentation for all data processing activities should ideally be in one location, but a single location for each process is sufficient. The steps need to be sufficiently documented so that anyone new to the project could use the documentation to follow and implement the processes.

To best implement a complete set of documentation, the suggested components of the metadata documentation (Section 4.3) should be followed.

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

**Criterion 23** *Technical specifications for the data holding are maintained.*

The way systems and applications are developed for a data holding can affect its data quality. Therefore, it is important that the specifications be implemented as intended and fully tested. It is equally important to have these technical specifications maintained for the data holding in the form of documentation. The reasons for documentation are simple: documentation allows easy validation of the systems, programs and applications, 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. Furthermore, such documentation should be regularly reviewed, and any updates to the technical specifications should also be reflected in the documentation.

This criterion is met if documentation about the data holding's systems, programs or applications are maintained.

**Criterion 24** *Changes to a data holding's underlying structure or processing or estimation programs have been tested.*

Although revisions are occasionally necessary to accommodate modifications to a data holding, changes to programs can have unexpected consequences. For instance, changes to data elements can require extensive testing, whether it is just a change to the data element name or an addition or deletion of a data element. The modified programs should be checked to ensure that expected results are achieved. It is also prudent to verify the downstream effects of the changes. In an example of the latter, changing the format of a data element from numeric to character can affect programs that later treat the data element as numeric. Unit, system and user-acceptance testing should be performed to prevent unexpected results in a production environment.

This criterion is met if the systems are tested when changes are made. If no revisions have taken place in the last year, then this rating is not applicable.

**Criterion 25** *Raw data, according to the CIHI policy for data retention, is 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 providers, be saved in a secure location. The data should be saved in such a way that it cannot be modified or deleted by accident. Having the unmodified data allows database personnel to refer to the original data. If the CIHI policy for data retention is followed, it will be far easier to handle situations that arise. For instance, errors that have occurred during data processing can be located, quantified and finally minimized through a subsequent run using the raw data. In addition, disputes that have arisen about the data can be resolved by going back to the raw data. As well, if the results of an analysis are questionable, the raw data can be very useful in verifying the original analysis by repeating the processing and analysis steps.

It should be noted that if the raw data cannot be saved or was not saved, the program area should, at a minimum, be able to recover data from the previous stages of data processing. This will, at least, permit the program area to rerun the processing steps.

This criterion is met if the data from data providers is saved in a secure location or if any changes to data can be reproduced.

**NEW**

**Criterion 26a** *Aggregated statistics from a data holding have been compared, where possible, to similar statistics from another CIHI data holding or external source.*

As a further indicator of good data quality, it is a worthwhile practice to compare statistics at the aggregate level from one data holding to another. This comparison should be done when aggregated statistics have been produced for the same data element within the same population of reference and time period. Where differences are noted, it may signal a problem with the aggregation process. Aggregated statistics from data holdings whose data is based on a census and those whose data is based on probability samples can be compared.

This criterion is met if the data holding, where possible, has compared its aggregated statistics with those from similar data holdings.

**NEW**

**Criterion 26b** *The variance of the estimate, compared to the estimate itself, is at an acceptable level.*

This criterion applies only to estimates that are based on a random sample for which the probability of selection is known. **Databases that do not use such samples should rate this criterion as *not applicable*.** In order to avoid confusing terms, it is important for those databases that are based on census data not to use the term variance when describing variability in their data, but rather use other measures of variability such as range, inter-quartile range, an average deviation or a mean absolute deviation.

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. An acceptable level for the variance will depend on the value of the estimate resulting from the sample. When making statements on an estimate's variability, it is important to report the standard error and/or coefficient of variation. The coefficient of variation (CV) is the standard error divided by the value of the estimate and expressed in percentage terms. This is the statistic that is the basis for the rating guideline table that follows. This table is to guide those releasing estimates based on sample data. Depending on the value of the CV for an estimate, this rating guideline can be used to determine if an estimate is acceptable to be published, requires a warning note to the user or indicates that it should not be published at all because of unreliability.

**Table D Rating Guide for Estimated CVs**

CV (Percent)	Guideline
Less than 16.6%	Acceptable level
16.6% to 33.3%	Proceed with caution
More than 33.3%	Do not publish

Table D is to be used as a guideline in determining if the sampling error is at an acceptable level or not. Take note that the CV that will be obtained depends greatly on the value of the estimate, since the estimate is the denominator in the calculation of the CV. In addition, the sample size with respect to the number of frame units in the population can also affect the variance. In general, CVs of estimates that reach an acceptable level (less than 16.6%) should rate this criterion as *met*.

The criterion should be rated as *not met* when the CV is above 16.6% or if the level is not acceptable to the clients of the data holding. This being said, the final decision on whether a variance is acceptable lies with the user and specific needs for the data.

This criterion is met if the variance of the estimate compared to the estimate itself is at a level that is acceptable to the data users.

## 2. Timeliness Dimension

Timeliness refers primarily to how up to date the data is at the time of release. The currency of the data is measured in terms of the gap between the end of the reference period to which the data pertains and the date on which the data becomes 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. Databases 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 hospital or private clinic staff to complete the survey on their medical imaging equipment, statistics on such equipment might be timely but incomplete or inaccurate. It might be argued that this sacrifice in completeness or accuracy is not worth the gain in timeliness. Sufficient time must also be set aside after the 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. Refer to the hierarchical view of the dimensions in Section 3.3 to further understand how timeliness relates to the other dimensions of data quality.

The purpose of the timeliness dimension is to examine how current the data is and whether the recommended data quality documentation was made available in time for release. Timeliness should be monitored to identify good practices at CIHI as well as to identify cases of extreme tardiness. The criteria in this dimension also assess whether major data holding reports are released on schedule. The dimension is composed of the following characteristics:

- Data currency at the time of release (is 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

## Data Currency at the Time of Release

### Criteria

- 27 *The difference between the actual date of data release and the end of the reference period is reasonably brief.*
- 28 *The official date of data release was announced before the release.*
- 29 *The official date of data release was met.*
- 30 *Data processing activities are regularly reviewed to improve timeliness.*

This characteristic first and foremost helps determine how current, or up to date, the data within a data holding is 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 relates. The duration should be short enough so that the data remains relevant for its main purposes. Also pertinent to data currency is whether the data is released on time and whether the data holding 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 data release and the end of the reference period is reasonably brief.*

The **date of release** is defined as the official date upon which an annual release of data from a data holding becomes available to users. The **reference period** refers to the period of time that the data actually spans or to which it relates. The start of the reference period is the first date to which the data relates, and the end of the reference period is the last date to which the data relates. For data holdings that do not have an annual release of data, a release of data to significant stakeholder(s) should be used as the point of comparison. For those databases that are longitudinal in nature and/or have no end of the reference period, this criterion is not applicable.

Data holdings will have different standards related to what is reasonably brief. As a general rule for annual data releases, a 6- to 9-month period between the end of the reference period and the release date is desirable, with up to 12 months being acceptable.

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 data release was announced before the release.*

Releases to a significant stakeholder(s), such as an annual release 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. Examples where such announcements can be made include the CIHI Calendar of Deliverables, the CIHI *Products and Services Guide*, CIHI's website and memos to key stakeholders. This is an important consideration for users, because it enables them in turn to 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 release or releases of data to a significant stakeholder(s) was planned for and announced at least six months in advance.

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

It is important to users that a release of data occur on time and as planned. The timing of the actual release date in relation to the planned release date may affect the production cycle of those who are dependent on the data. Monitoring the achievement of pre-announced release dates, changes to the release dates and reasons for changes is recommended. In order to reach earlier release dates without significantly affecting the other dimensions of quality, past experience should be used to adjust the timing of the release date, and product areas should always strive to shorten the production cycle. The production cycle could possibly be shortened by developing efficient methods to process the data more quickly.

This criterion is met if the data was released on or before the official date of data release. For data releases to a significant stakeholder(s), this criterion is to be based on the planned release date (as reported in the Calendar of Deliverables) versus the actual release date.

**Criterion 30** *Data processing activities are regularly reviewed to improve timeliness.*

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 to produce timely data. 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 data holding methods to similar external or internal holdings may yield insights that result in improved efficiency. Existing methods may be reviewed in light of new technologies, procedures or standardized practices across data holdings that might be more efficient (such as eDSS, or electronic Data Submission Services) and may result in more accurate data at the same time. New methods or technologies, as long as they have been thoroughly tested, may be an ideal way to improve timeliness and accuracy at the same time.

This criterion is met if data holding processing activities were reviewed at least once in the previous year.

## Documentation Currency

### Criteria

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

*32 Major 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 data holding 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 is 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 can internally or externally access the data or the reports summarizing the data. Therefore, a sufficient amount of time between the closure of the database and release of the data must be allocated for data quality documentation. For longitudinal databases, data quality documentation must be available whenever data is either accessed or released or when a report is written. Data quality documentation will include such information as the background, data sources and methodology, concepts and variables measured, data accuracy, data comparability, publications and products, and a list of services for further information. The recommended components of good data quality documentation are further discussed in Section 4.

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

**Criterion 32** *Major reports were released on schedule.*

It is important that the release dates for major reports be announced in advance and that the reports be released on schedule. Failure to meet deadlines with written reports not only inconveniences users who have planned activities around the scheduled release date, it can also undermine user confidence.

This criterion is met if the major reports for the data holding were released on schedule.

### 3. Comparability Dimension

Comparability is defined as the extent to which data holdings are consistent over time and use standard conventions (such as data elements or reporting periods), making them similar to other data holdings. Within an organization like CIHI, with many different data holdings, 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. Same or similar populations of reference are necessary when making comparisons between data holdings. 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 different 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 (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 databases 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 composed 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 (is 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



## Data Dictionary Standards

### Criteria

*33 All data elements are evaluated to determine their inclusion within the CIHI Data Dictionary.*

*34 Data elements from a data holding that are included within the CIHI Data Dictionary must conform to dictionary standards.*

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 in the process of being reviewed and revised. At the time of this latest revision of the Data Quality Framework, all the data elements under the object class of “Health Care Provider” have been finalized. Please refer to [www.cihi.ca/datadictionary](http://www.cihi.ca/datadictionary) for the latest definitions of the finalized data elements. While the adoption of the data dictionary standards is mandatory for newly developed data holdings (or those being redeveloped), the standards do not currently have to be applied to existing data holdings; however, it is recommended to ensure comparability across the holdings.

**Criterion 33** *All data elements are evaluated to determine their inclusion within the CIHI Data Dictionary.*

The CIHI Data Dictionary is the standard for data elements that all data holdings at CIHI should follow. As the dictionary continues to evolve, many of the data elements pertaining to health care providers have updated definitions in the CIHI Data Dictionary. Consequently, it is important that all data elements in existing data holdings and those being developed be reviewed periodically against the CIHI Data Dictionary (see [www.cihi.ca/datadictionary](http://www.cihi.ca/datadictionary)). The evaluation will determine if the standards used for the data elements in a data holding are in agreement with those in the CIHI Data Dictionary. Those that do not agree with the data dictionary standard should be noted and considered for future modification. In addition, it should be noted which data elements from the holding do not currently have a standard in the CIHI Data Dictionary.

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

**Criterion 34** *Data elements from a data holding that are contained within the CIHI Data Dictionary must conform to dictionary standards.*

Any data elements in a data holding that are also common to finalized data elements in the CIHI Data Dictionary (see [www.cihi.ca/datadictionary](http://www.cihi.ca/datadictionary)) 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, all attributes should be the same. When they are not, conformance may be partial or not exist at all. For example, a data element representing the units by which a health care provider’s gender is measured should have the following attributes:

Attribute	Attribute Description
Name	Provider’s Gender
Data Type	Character
Maximum Length	1
Value Domain	M = male F = female U = undifferentiated 7 = not collected 8 = not applicable 9 = unknown

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 value domains are more serious, since every record needs to be changed in order to conform to a given standard.

**Note:** The CIHI Data Dictionary is not currently complete for all types of data elements; therefore, the assessment of conformance only applies to those elements that are currently contained within it. Any justifiable deviations from CIHI standards should be described.

This criterion is met if all of the data elements common to the data holding and the CIHI Data Dictionary are an exact match.

## Standardization

### Criteria

*35 Data is collected at the finest level of detail practical.*

*36 For any derived data element, the original data element remains accessible.*

Databases often group data 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. It should be noted that the data dictionary does not have a standard for provider age but does have a standard for provider birthdate, birth year and birth month. Provider age can therefore be derived from any of these data dictionary standards.

**Criterion 35** *Data is collected at the finest level of detail practical.*

A fine level of detail in data definitions is important, because it allows flexibility to conform to different standards. It is important to note that the level of detail depends on the users of the data holding and, in some cases, the data provider. 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. Another example relates to the collection of data by using a person's birthdate rather than capturing age groups, even if the information is reported in certain age group categories. A simple regrouping into different age ranges is possible when age has been reported in years. In summary, 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 all core data elements are collected with the necessary detail required for publication and for linking or comparison purposes. Any exceptions to capturing the data at the finest detail need to be justified.

**Criterion 36** *For any derived data element, the original data element remains accessible.*

As a general rule, data elements used in the creation of another data element need to be maintained in the event that changes or new calculations have to be made. Sensitive items, such as health card number or birthdate, may be encrypted or 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. These sensitive items should not be permanently deleted from the raw file of the main database; rather, an additional data set containing just the derived data elements should be created for which all sensitive items have been removed or encrypted. 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 original data elements are accessible and are not permanently deleted.

## Linkage

### Criteria

- 37 Geographical data is collected using the Standard Geographical Classification (SGC).*
- 38 Data is collected using a consistent time frame, especially between and within jurisdictions.*
- 39 Identifiers are used to differentiate facilities or organizations uniquely for historical linkage.*
- 40 Identifiers are used to differentiate persons or machines uniquely for historical linkage.*

**Linkage** refers to the process of joining records from two or more data holdings by the use of one or more common linking data elements, or joining records within a holding through one or more common data elements. Given the variety of data at CIHI, linkage of data from different sources provides a more complete picture. In addition, data quality can be improved with linkage, as duplicates can be identified and removed. **Keep in mind that CIHI's privacy and confidentiality guidelines must be adhered to when linking data between data holdings. CIHI staff should refer to the Privacy and Legal Services Secretariat intranet page.**

In order to link data, one should follow a three-step process.

1. Prepare data—involves the removal of duplicates and prepping the data sets in the same format (SAS, Excel, etc.).
2. Choose linking data elements (one or more)—choose a data element to be a unique identifier that is present on all files being linked. All data elements used in the linking are to be very reliable and accurately recorded. The Service Recipient Index (SRI) should be used if available for data sets in question.
3. Evaluate each record being linked—either a success or failure (unique identifier not on all files or a link was made incorrectly due to typos, etc.).

**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 or machine.

**Criterion 37** *Geographical data is collected using the Standard Geographical Classification (SGC).*

This criterion relates to the **Standard Geographical Classification (SGC)** as defined by Statistics Canada. The SGC is a classification of geographical areas used to collect and disseminate statistics. Within it, codes of standard geographical areas (census tracts, census subdivisions, census divisions, census metropolitan areas, census agglomerations and economic regions) are organized by province and territory and by metropolitan area. 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 into province and country. Much social and demographic information derived from the census is aggregated at different levels of the SGC, making it valuable for a wide range of research purposes. For linkage purposes, it is important that geographical data collected by each data holding be in agreement with the SGC. For instance, the capture of the full postal code is sufficient since it can then be converted to the SGC by way of the **Postal Code Conversion File**.

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

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

**Criterion 38** *Data is collected using a consistent time frame, especially between and within jurisdictions.*

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 on fiscal year. It is important to collect data for a database using a consistent time frame over the years in order to avoid gaps in the database. Also, dates should be collected at the finest level of detail (YYYYMMDD). By following the standard format for dates, calendar year data results can be produced even if the data has previously been collected on a fiscal year basis.

This criterion is met if sufficient information between and within jurisdictions is available that would allow the data to be analyzed using a consistent time frame.

**Criterion 39** *Identifiers are used to differentiate facilities or organizations uniquely for historical linkage.*

Facilities or organizations are a common level of analysis at CIHI and should therefore have a unique identification code in order to facilitate historical record linkage. Usually, the province assigns its facilities or organizations a numeric code identifier. Other identifiers are acceptable if a list of changes is maintained or a suitable cross-reference table is available, such as CIHI's Organizational Index or the Institutional Care Facility Master Inventory (ICFMI) used by Statistics Canada. If numeric code identifiers are not unique between provinces, one possibility is to combine province code and institution number to create a unique identifier for linkage purposes. 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 can often occur across different time periods.

This criterion is met if a unique code acceptable for historical linkage purposes (provincially assigned identifier or equivalent) exists for each facility or organization and is available on the database.

**Criterion 40** *Identifiers are used to differentiate persons or machines uniquely for historical linkage.*

The purpose of this criterion is to ensure that a suitable identifier is present that accurately distinguishes between persons or machines (that is, medical devices) 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 or machines. 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 Service Recipient Index (SRI) or 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 health provider or personnel databases, other identifiers may be appropriate, such as those assigned by the province or regulatory body. It is important to note that names are not suitable for linking purposes because they are subject to typos, etc. and although they are unique for an individual, different individuals can have the same name. For data holdings that collect information on medical devices, identifiers such as serial numbers or some other unique machine identifier should be considered when a machine is composed of numerous parts that have their own serial numbers.

This criterion is met if a unique person or machine identifier is available in the database that could be used to link to corresponding records across different time periods.

## Equivalency

### Criteria

- 41 Methodology and limitations for crosswalks and/or conversions are documented.*
- 42 The magnitude of issues related to crosswalks and conversions falls into one of the predetermined categories.*

Equivalency refers to how well data can be mapped over time, especially when classification systems are included (for example, the International Classification of Diseases [ICD], Anatomical Therapeutic Chemical codes [ATC], MIS, etc.). Crosswalks and conversions are simply tables that are used to do such mapping. An example of equivalency relevant to CIHI includes the mapping required as one relates data under one ICD to another (for example, ICD-10-CA to ICD-9).

In the case of **conversions**, the mapping is one-to-one. For example, one could create a table that maps a patient's weight in pounds to kilograms. In this case, the conversion is simple, since the formula is straightforward.

**Crosswalks** involve a many-to-one or one-to-many relationship. Crosswalks are involved in the mapping of diagnosis codes. For instance, a patient’s chart containing diagnosis codes in ICD-9-CM can have the codes mapped to ICD-10-CA. In such a situation, one code in ICD-9-CM could be mapped to several codes in ICD-10-CA or several codes in ICD-9-CM could be mapped to one code in ICD-10-CA.

The success of a crosswalk or conversion is based largely on how well it can convert values from one classification to another.

**Criterion 41** *Methodology and limitations for crosswalks and/or conversions are documented.*

Due to the complexity of many crosswalks or conversions, the methodology and limitations need to be adequately documented on an annual basis at least. Any enhancements or alterations should be explained. Information derived by crosswalks or conversions also needs to be documented in reports as such. For databases that have data extractions occurring at any time, it is essential to always have the documentation on the crosswalks and conversions up to date. Finally, 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 at least annually.

**Criterion 42** *The magnitude of issues 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	No or few issues
Moderate	Identifiable issues that are limited in scope
Significant	A significant portion of the data is not being converted properly, and this has an impact on results
Unknown	Equivalency has not been investigated



## Historical Comparability

### Criteria

- 43 *Documentation on historical changes to the data holding exists and is easily accessible.*
- 44 *Trend analysis is used to examine changes in core data elements over time.*
- 45 *The magnitude of issues associated with comparing data over time falls into one of the predetermined categories.*

Historical comparability refers to the consistency of data concepts and methods over time, which in turn allows valid comparisons of different estimates at different points in time. Many things can make the comparison of data over time difficult. Database enhancements that will improve a database for the future can sometimes inhibit historical comparability. In those situations, this limitation should be noted in data user documentation.

**Criterion 43** *Documentation on historical changes to the data holding exists and is easily accessible.*

This criterion assesses whether documentation of historical changes exists and is maintained in one document. It should include changes to concepts, methodologies, frames 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. In addition, storing comments within a SAS program is also not an acceptable form when it is the sole source of historical documentation. Major changes from previous years should be included in the data quality documentation, but a more detailed document for internal use may also be necessary. The metadata documentation is the recommended source to hold such information for internal use; refer to Section 4.3 for more details.

This criterion is met if a single document of historical changes is maintained and updated on an annual basis.

**Criterion 44** *Trend analysis is used to examine changes in core data elements over time.*

Trend analysis includes comparisons of counts or proportions over time, as well as more sophisticated time series analysis, smoothing or curve fitting. Graphing data is often particularly helpful for investigating temporal changes. Within a physician database, one might examine changes in the number of physicians for a particular geographical area 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 is expected to naturally trend upward or downward due to policies implemented or social or economic changes.

This criterion is met if trend analysis has been performed for core data elements since the last data quality assessment.

**Criterion 45** *The magnitude of issues associated with 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 taken into account. 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 general guide for assessing this criterion.

<b>Suggested Rating</b>	<b>Guideline</b>
Minimal	No or few issues in producing comparable trends
Moderate	Issues 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

## 4. Usability Dimension

Usability reflects the ease with which data from a data holding 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 data holding'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 improve the standardization of data benefit not only the ease with which the data can be used, but also the accuracy of the data. Inconsistent 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 the time data is released and on an ongoing basis after release. Once major limitations are known, they should be documented for users. Data holding methods and changes to the methods should also be documented for users. Also, the data has to be in a readily accessible user-friendly form. Finally, no matter how well documented or accessible, if users are not aware of the existence of a data holding, the data will not be used.

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

- Accessibility (how readily accessible is the data?)
- Documentation (how well documented is 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

## Accessibility

### Criteria

- 46 *A final data set is made available per planned release.*
- 47 *Standard tables and analyses using standard format and content are produced per planned release or upon request.*
- 48 *Products are defined, catalogued and/or publicized.*

This characteristic deals with the ease with which data from a data holding can be obtained from CIHI. This includes the ease with which the existence of the holding 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 for whatever reason will not be of use to them.

#### **Criterion 46** *A final data set is made available per planned release.*

The data that is 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. Where analyses are done directly on a database, it is important to have the SAS codes used to produce the tables maintained for one year after the release of a report.

This criterion is met if the data file from a data holding (or SAS code where a data set is not possible) is made available to stakeholders following a planned release.

#### **Criterion 47** *Standard tables and analyses using standard format and content are produced per planned release or upon request.*

For many users, aggregate statistics or summary tables are more useful than microdata. In addition to major reports, aggregate statistics and standard tables should also be made available for users, per planned 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 or released to confirm that they are reasonable. For some users, the microdata file may contain the information that is the most beneficial. In such cases, past formats of the microdata should be followed as much as possible to allow for comparisons to be made.

This criterion is met if commonly used standard tables and analyses are made available per planned release or upon request.

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

To assist users, a data holding's associated products (annual reports, analytical reports, custom reports, etc.) should be listed in the corporate-wide dissemination systems. These include the CIHI website, the CIHI Calendar of Deliverables and the CIHI *Products and Services Guide*. The website can be used as a virtual library of all the information products that are available for the public from CIHI. Through the website, requests can be made through the Graduate Student Data Access Program (GSDAP), Special Needs and Application Program (SNAP) or special research and raw data request options. Other channels, such as the press (via media releases), research data centres and public libraries may also be used to gather information on CIHI data holdings.

This criterion is met if the products of a data holding are defined, catalogued and/or publicized in any of the CIHI dissemination systems per planned release.

**Documentation****Criteria**

49 *Current data quality documentation for users exists.*

50 *Current metadata documentation exists.*

51 *A caveat accompanies any 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 data from a data holding. Documentation will include such information as the background, data sources and methodology, concepts and data elements measured, data accuracy and data comparability (see Section 4 for more details). Throughout the entire data quality work cycle (mentioned in Section 2), documentation should be done on an ongoing basis during the various planning, implementation and assessment phases. What has been measured, how it was measured and how well it was measured needs to be clearly documented for users.

**Criterion 49** *Current data quality documentation for users exists.*

The purpose of data quality documentation for users is to give both internal and external users of the data holding sufficient information so they can decide if the quality of the data is appropriate for their intended use. Contact information should also be provided with any release so that users can access additional information on the limitations, which they may require for their intended use.

A stand-alone data quality document for users should be made available at least once a year. See Section 4.2 for more information on the data quality documentation for users.

This criterion is met if updated data quality documentation for users exists any time data is released or extracted.



**Criterion 50** *Current metadata documentation exists.*

To facilitate the interpretation and proper use of the data, all data holding methods documentation should be made readily available to CIHI staff working with the data. This documentation is referred to as the “metadata documentation” in Section 4.3. While data quality documentation for users provides some background information and outlines the major data limitations, detailed background notes, as well as all known data limitations, etc. should be made available internally through the metadata documentation.

This criterion is met if metadata documentation for the data holding exists for internal purposes. This documentation should be reviewed on an annual basis and updated as required.

**Criterion 51** *A caveat accompanies any preliminary release.*

In an effort to improve timeliness, some data holdings may provide preliminary releases of data or results. If the data holding does not release any preliminary data or report based on preliminary data, this criterion is not applicable.

A **preliminary release** may include a release 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 or ministries for verification and validation. A **preliminary release** may also be defined as a release of possibly incomplete data for the purpose of improved timeliness. For example, data that is collected on an annual basis might be released six months prior to year-end so that health care system planners can gain an early indication of the complete data to follow.

For all preliminary releases, a caveat must be provided that advises that the data may not be complete and that it is subject to revision. A description of both the unit and item response rates to date, as well as the expected final unit and item response rates, should be included. The anticipated revisions and their possible impact should also be conveyed.

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

## Interpretability

### Criteria

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

*53 Revision guidelines are available and applied per 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 generalizations made with respect to 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. Having the record layouts accompanying the various data files will also aid in the interpretability of the data.

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 data holding program area.*

Contact information (name, phone number and email address) should be included with releases so that users (internal or external) can provide feedback on any major data quality limitations as they come to light. Major users should be encouraged to use the contact information to provide feedback on any limitations they may discover or concerns that they may have.

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 comprising key users (for example, Statistics Canada, Health Canada and CIHI analysts).

This criterion is met if data holding program area contact information is included with any major data or report release and feedback is solicited.

**Criterion 53** *Revision guidelines are available and applied per release.*

Initial estimates can be revised as errors or missing data come to light. As the resolution of any errors and/or previously missing data becomes available, a revision may be deemed necessary. A **revision** is a change to the data, or to the estimates based on the data, once the data has 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 would be released.

This criterion is met if data holding-specific revision guidelines are available and applied whenever data is released or extracted. These guidelines need to be current and in documented form.

## 5. Relevance Dimension

Relevance reflects the degree to which a data holding 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 informs the issues most important to users. In addition to ensuring that its data is 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 data holding can adapt to change and whether the holding is perceived to be valuable. Relevance is composed of the following characteristics:

- Adaptability (can user needs be anticipated and planned for?)
- Value (how valuable is the data?)

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

### Adaptability

Criteria
<i>54 Mechanisms are in place to keep stakeholders informed of developments in the field.</i>
<i>55 The data holding is developed so that future system modifications can be made easily.</i>

The adaptability of a data holding 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 data holding 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 something is anticipated, future information needs can be factored into the design of the data holding. Although it is impossible to predict the future needs of users with complete accuracy, one can try to design data holdings that allow for change. It is also important that a data holding not be overly flexible to the wants and needs of users. The impact of changes required on the entire work cycle should be fully evaluated before being incorporated.



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

Maintaining a stakeholder liaison by the data holding program 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. Program 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. Online query systems (where applicable) can also be used as a mechanism for clients or stakeholders to remain informed. An example of this is the CIHI online coding query database that helps answer users' coding questions. If main users or stakeholders are common across holdings, then a coordinated approach to the consultative process may be considered across data holdings.

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

**Criterion 55** *The data holding is developed so that future system modifications can be made easily.*

In order for a data holding 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 collect emergency room wait times in an ambulatory care database. A data holding should also be able to incorporate new technical standards as they arise (for example, ICD-10-CA).

In addition to dealing with important emerging issues or with new technical standards, changes to a data holding 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 data holding area's control might all be considered part of ongoing improvement.

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

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

## 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 assessed.*

The value of a data holding may be defined by its contribution to population health or health care system knowledge and to its use. That is, the worth or utility of a holding 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 data holding's data should also be monitored. 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 data holding depends on whether it fills a health information gap. The mandate of the holding should be periodically assessed in relation to the other data holdings within CIHI and across the field externally. How a data holding 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 data holding may be related to the extent the data is used. Evidence of usage may include high-profile uses of the data (for example, the Romanow Report), web page hits, press clippings, news items, citations, staff-authored papers, sales, media appearances/contacts of staff, conferences and policy forums. As well, the level of usage can be monitored by keeping track of the number and type of data requests and, where possible, the use of eReports.

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

**Criterion 58** *User satisfaction is periodically assessed.*

It is important to assess whether the data holding 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 a data holding's perceived value. A satisfaction survey may 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 query is used or a client support hotline is in place. A satisfaction survey may also reveal that having a database based on mandated submissions has significantly more value than one that is built on voluntary data submissions. Due to resource constraints, such a revelation may lead to having a sample selected from the population of reference and expecting those selected to be mandated submitters. In any event, the impact of any suggested modifications resulting from a stakeholder satisfaction survey needs to be fully assessed prior to any kind of implementation.

This criterion is met if client satisfaction assessments are conducted periodically.





## Executive Summary

An executive summary should be included to summarize the strong points of the data holding and highlight problem areas. A tabular view of results may also be used as shown below.

### Summary of Criteria Assessment

Dimensions/ Characteristics	Criteria	Assessment
<b>Accuracy</b>		
<b>Coverage</b>	1a The population of reference is explicitly stated in all releases.	
	1b Efforts are being made to close the gap between the population of reference and the population of interest.	
	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>	5a CIHI practices that minimize response burden are documented.	
	5b CIHI has documentation of data-provider practices that minimize response burden.	
	6 Practices exist that encourage cooperation for data submission.	
	7 Practices exist that give support to data providers.	
	8 Standard data submission procedures exist and are followed by data providers.	
	9 Data-capture quality control measures exist and are implemented by data providers.	
<b>Unit Non-Response</b>	10 The magnitude of unit non-response is mentioned in the data quality documentation.	
	11 The number of records for responding units is monitored to detect 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>Dimensions/ Characteristics</b>	<b>Criteria</b>	<b>Assessment</b>
<b>Edit and Imputation</b>	18 Validity checks are done for each data element and any invalid data is flagged.	
	19 Edit rules and imputation are logical and applied consistently.	
	20 Edit reports for users are easy to use and understand.	
	21 The imputation process is automated and consistent with the edit rules.	
<b>Processing and Estimation</b>	22 Documentation for all data processing activities is maintained.	
	23 Technical specifications for the data holding are maintained.	
	24 Changes to a data holding's underlying databases or processing or estimation programs have been tested.	
	25 Raw data, according to the CIHI policy for data retention, is saved in a secure location.	
	26a Aggregated statistics from a data holding have been compared, where possible, to similar statistics from another CIHI data holding or external source.	
	26b The variance of the estimate, compared to the estimate itself, is at an acceptable level.	
<b>Timeliness</b>		
<b>Data Currency at the Time of Release</b>	27 The difference between the actual date of data release and the end of the reference period is reasonably brief.	
	28 The official date of data release was announced before the release.	
	29 The official date of data release was met.	
	30 Data processing activities are regularly reviewed to improve timeliness.	
<b>Documentation Currency</b>	31 The recommended data quality documentation was available at the time of data or report release.	
	32 Major data holding reports were released on schedule.	
<b>Comparability</b>		
<b>Data Dictionary Standards</b>	33 All data elements are evaluated to determine their inclusion within the CIHI Data Dictionary.	
	34 Data elements from a data holding that are contained within the CIHI Data Dictionary must conform to dictionary standards.	
<b>Standardization</b>	35 Data is collected at the finest level of detail practical.	
	36 For any derived data element, the original data element remains accessible.	

<b>Dimensions/ Characteristics</b>	<b>Criteria</b>	<b>Assessment</b>
<b>Linkage</b>	37 Geographical data is collected using the Standard Geographical Classification (SGC).	
	38 Data is collected using a consistent time frame, especially between and within jurisdictions.	
	39 Identifiers are used to differentiate facilities or organizations uniquely for historical linkage.	
	40 Identifiers are used to differentiate persons or machines uniquely for historical linkage.	
<b>Equivalency</b>	41 Methodology and limitations of crosswalks or conversions are documented.	
	42 The magnitude of issues related to crosswalks or conversions falls into one of the predetermined categories.	
<b>Historical Comparability</b>	43 Documentation on historical changes to the data holding exists and is easily accessible.	
	44 Trend analysis is used to examine changes in core data elements over time.	
	45 The magnitude of issues associated with comparing data over time falls into one of the predetermined categories.	
<b>Usability</b>		
<b>Accessibility</b>	46 A final data set is made available per planned release.	
	47 Standard tables and analyses using standard format and content are produced per planned release or upon request.	
	48 Products are defined, catalogued and/or publicized.	
<b>Documentation</b>	49 Current data quality documentation for users exists.	
	50 Current metadata documentation exists.	
	51 A caveat accompanies any 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 release.	
<b>Relevance</b>		
<b>Adaptability</b>	54 Mechanisms are in place to keep client and stakeholders informed of developments in the field.	
	55 The data holding is developed so that future system modifications can be made easily.	
<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 assessed.	



## Recommendations

Recommendations that are made throughout the report must be summarized in an action plan in the executive summary. The action plan identifies the recommendation, when the work will be initiated and completed, who will be responsible and whether the Data Quality department is required. Below is a blank template to be completed for the action plan.

### Action Plan

**NEW**

Recommendation	When Initiated?	Staff Responsible	Target Date or Ongoing	DQ Involvement?
1.				
2.				
3.				
4.				
5.				

## Introduction

It is suggested that the report start with a brief introduction describing the data holding and the time frame for the assessment.

## Detailed Assessment

The dimensions, characteristics and criteria should be reported here. The ratings for each criterion should be supported by at least one or two sentences describing why the rating was given and, when applicable, documentation supporting the rating should be included or at least referenced.

## 1. Accuracy Dimension

### 1.1 Coverage

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

Your assessment of the criterion goes here.

**Details Required:** First, provide details on the following for the data holding: population of interest, population of reference, frame units, units of analysis and reference period. Second, indicate whether the population of reference is explicitly stated in all releases. Include the title of the release document(s) and where it (they) may be viewed. If the population of reference as stated in the release(s) varies from year to year or document to document, provide details about this as well.

**Assessment:**       Met       Not met



**Criterion 1b** *Efforts are being made to close the gap between the population of reference and the population of interest.*

Your assessment of the criterion goes here.

**Details Required:** If any difference between the population of reference and the population of interest has been discussed, provide details about any efforts to close the gap between these two populations.

**Assessment:**       Met       Not met       Unknown       Not applicable

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

Your assessment of the criterion goes here.

**Details Required:** Provide the sources and reasons for under- and over-coverage. Note that each level of observation (that is, frame unit) should be mentioned as described in Table B of the Accuracy Dimension section of Appendix C.

**Assessment:**       Met       Not met       Unknown

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

Your assessment of the criterion goes here.

**Details Required:** Report on efforts to validate the frame through comparison to sources external to and independent of the data holding. Provide the details regarding sources used and their credibility. Note that the use of multiple sources and comparison at an aggregate level may also be acceptable.

**Assessment:**       Met       Not met       Not applicable

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

Your assessment of the criterion goes here.

**Details Required:** Provide the calculations and any supporting definitions for the data holding rates of under- and over-coverage. Note that the assessment is driven by the result of the calculations.

- Assessment:**
- None or minimal (less than 1%)
  - Moderate (1% to 5%)
  - Significant (greater than 5%)
  - Unknown (could not be determined)

## 1.2 Capture and Collection

**Criterion 5a** *CIHI practices that minimize response burden are documented.*

Your assessment of the criterion goes here.

**Details Required:** Provide information on the CIHI practices employed to minimize response burden, as well as the name of and the location where documentation of these practices may be accessed.

- Assessment:**
- Met
  - Not met
  - Unknown

**NEW**

**Criterion 5b** *CIHI has documentation about data-provider practices that minimize response burden.*

Your assessment of the criterion goes here.

**Details Required:** Provide information on the practices followed by data providers to minimize response burden, as well as the name of and location where documentation of these practices may be accessed.

- Assessment:**
- Met
  - Not met
  - Unknown

**Criterion 6** *Practices exist that encourage cooperation for data submission.*

Your assessment of the criterion goes here.

**Details Required:** Note the nature of participation for data providers (that is, mandated, voluntary) and the practices employed by CIHI to encourage cooperation.

- Assessment:**
- Met
  - Not met
  - Unknown

**Criterion 7** *Practices exist that give support to data providers.*

Your assessment of the criterion goes here.

**Details Required:** Supply details regarding data holding practices that give support to data providers before and during data capture, for example, education sessions, prompt response to emails and phone calls, technical and coding support, access to supporting documentation, coding guidelines and the abstracting manual.

**Assessment:**         Met         Not met         Unknown

**Criterion 8** *Standard data submission procedures exist and are followed by data providers.*

Your assessment of the criterion goes here.

**Details Required:** Information regarding standard data submission forms and procedures for the data holding, as well as the extent to which these are followed by data providers.

**Assessment:**         Met         Not met         Unknown

**Criterion 9** *Data-capture quality control measures exist and are implemented by data providers.*

Your assessment of the criterion goes here.

**Details Required:** Note measures taken to ensure that data is recorded properly. Include information on whether these measures are carried out by data holding staff or data providers. It may be appropriate to mention the ability of data providers to maintain such measures, as well as any work done by the data holding to enhance data capture at the provider level.

**Assessment:**         Met         Not met         Unknown

### 1.3 Unit Non-Response

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

Your assessment of the criterion goes here.

**Details Required:** Report on whether the magnitude of the frame unit non-response is reported in the data quality documentation provided to users. (The frame unit could be the patient, service provider, health region, province or other data provider.) In addition, note where and how this documentation can be accessed, both by users and internally.

**Assessment:**         Met         Not met

**Criterion 11** *The number of records for responding units is monitored to detect unusual values.*

Your assessment of the criterion goes here.

**Details Required:** Provide information about program area procedures to track the number of units responding over time. Note existing challenges to these procedures and efforts to address them.

**Assessment:**  Met  Not met

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

Your assessment of the criterion goes here.

**Details Required:** Provide the calculations and any supporting definitions for the data holding rates of unit non-response. As this criterion considers both the frame level and the unit-of-analysis level, it is suggested that the non-response be calculated for each level of observation. A unit non-response rate is usually computed at CIHI rather than its complement, the unit response rate. Note that the assessment is driven by the result of the calculations.

**Assessment:**  None or minimal (non-response rate less than 2%)  
 Moderate (2% to 10%)  
 Significant (greater than 10%)  
 Unknown (could not be determined)

## 1.4 Item (Partial) Non-Response

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

Your assessment of the criterion goes here.

**Details Required:** Identify the processes and programs used by the data holding to distinguish between blank values and non-response for core data elements. To meet this criterion, the name and location of the document(s) where item non-response is identified needs to be specified.

**Assessment:**  Met  Not met

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

Your assessment of the criterion goes here.

**Details Required:** Provide the calculations and any supporting definitions for the level of non-response for each core data element. An item non-response rate is usually computed at CIHI rather than its complement, the item response rate. When providing the assessment for this criterion, the core data element with the highest item non-response rate should be considered.

- Assessment:**
- None or minimal (item non-response rate less than 2%)
  - Moderate (2% to 10%)
  - Significant (greater than 10%)
  - Unknown (could not be determined)

## 1.5 Measurement Error

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

Your assessment of the criterion goes here.

**Details Required:** Provide the calculations and any supporting definitions for the level of error rate for non-subjective variables. In assessing this criterion, the rating for the core data element with the highest error rate should be used. The amount of error in the data elements of a database is most often assessed through reabstraction or other special (and usually retrospective) studies, but it can also be assessed when the database is being developed. If the level of error is not estimated through a data quality study, it does not necessarily mean the criterion is automatically rated as *unknown*. See Appendix C for more details.

- Assessment:**
- None or minimal (error rate 0% to less than 5%)
  - Moderate (5% to 10%)
  - Significant (greater than 10%)
  - Unknown (could not be determined)

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

Your assessment of the criterion goes here.

**Details Required:** Provide information on whether there is, or is perceived to be, a substantial bias in the data. 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*.

**Assessment:**            Met            Not met            Unknown

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

Your assessment of the criterion goes here.

**Details Required:** If the database has data elements that depend on the opinion or interpretation of the coders, provide information on both the consistency of the measurements from the individual coders and the consistency of measurements between coders. When providing the assessment for this criterion, the rating for data elements with the lowest level of consistency should be used.

**Assessment:**            None or minimal (discrepancy rate\* 0% to less than 5%)  
 Moderate (5% to 10%)  
 Significant (greater than 10%)  
 Unknown (could not be determined)

\* Kappa statistic may also be used; see Appendix C for details.

## 1.6 Edit and Imputation

**Criterion 18** *Validity checks are done for each data element and any invalid data is flagged.*

Your assessment of the criterion goes here.

**Details Required:** Indicate whether validity checks are done for each data element, as well as the type of check carried out, that is, comparing the response to a list of acceptable responses versus ensuring that the response is in a proper format. Note how invalid data can be identified and any procedures for follow-up (invalid data excluded, sent back to the supplier for correction, flagged for imputation or flagged as invalid and dealt with separately).

**Assessment:**            Met            Not met

**Criterion 19** *Edit rules and imputation are logical and applied consistently.*

Your assessment of the criterion goes here.

**Details Required:** Provide details on the edit rules and imputation performed on the data, including the stage at which they are applied (data capture, data processing).

**Assessment:**       Met       Not met

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

Your assessment of the criterion goes here.

**Details Required:** Describe the data holding approach to generating edit reports for users, the clarity of reporting and the actions requested from recipients of the edit reports. In addition, any feedback from users on edit reports would be helpful here.

**Assessment:**       Met       Not met       Not applicable

**Criterion 21** *The imputation process is automated and consistent with the edit rules.*

Your assessment of the criterion goes here.

**Details Required:** Report on the imputation process, if any, followed by the data holding. Note whether imputation is automated and driven by edit rules.

**Assessment:**       Met       Not met       Not applicable

## 1.7 Processing and Estimation

**Criterion 22** *Documentation for all data processing activities is maintained.*

Your assessment of the criterion goes here.

**Details Required:** Provide information regarding all processing activities run by data holding personnel, how they are documented and where the documentation may be accessed.

**Assessment:**       Met       Not met

**Criterion 23** *Technical specifications for the data holding are maintained.*

Your assessment of the criterion goes here.

**Details Required:** Provide information regarding the data holding's systems, programs and/or applications, how they are documented and where the documentation may be accessed.

**Assessment:**       Met       Not met



**Criterion 24** *Changes to a data holding’s underlying structure or processing or estimation programs have been tested.*

Your assessment of the criterion goes here.

**Details Required:** Note any changes to the data holding’s underlying databases or processing or estimation programs for the reference period, as well as details on the type of testing (unit, system and/or user-acceptance testing) carried out. If no revisions have taken place in the last year, then this rating is not applicable.

**Assessment:**       Met       Not met       Not applicable

**Criterion 25** *Raw data, according to the CIHI policy for data retention, is saved in a secure location.*

Your assessment of the criterion goes here.

**Details Required:** Provide information regarding the location and storage procedures for the data holding’s raw data.

**Assessment:**       Met       Not met

**NEW**

**Criterion 26a** *Aggregated statistics from a data holding have been compared, where possible, to similar statistics from another CIHI data holding or external source.*

Your assessment of the criterion goes here.

**Details Required:** Provide details, where possible, of the data holdings whose aggregate statistics were compared and describe the results. For comparisons that did not result in an exact match, explanations should be provided. Pay special attention to discrepancies resulting from any differences in reference period, population of reference, source of data and timing of data collection.

**Assessment:**       Met       Not met       Not applicable

**Criterion 26b** *The variance of the estimate, compared to the estimate itself, is at an acceptable level.*

Your assessment of the criterion goes here.

**Details Required:** Provide information on the value calculated for the coefficient of variance (CV), as well as the acceptable level for major clients of the data holding. 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*. It is important that those databases that are based on census data do not use the term variance when describing variability in their data, but rather use other measures of variability such as range, inter-quartile range, an average deviation or a mean absolute deviation.

- Assessment:**
- Met (CVs of estimates that reach an acceptable level, less than 16.6%)
  - Not met (CV is 16.6% or greater or at a level that is not acceptable to the major clients of the data holding)
  - Not applicable

## 2. Timeliness Dimension

### 2.1 Data Currency at the Time of Release

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

Your assessment of the criterion goes here.

**Details Required:** State the difference between the actual date of data release and the end of the reference period. Provide information regarding any delays and whether the delays are one-time events or expected to continue. For data holdings that do not have an annual release of data, a release of data to significant stakeholder(s) should be used as the point of comparison. For those databases that are longitudinal in nature and/or have no end for the reference period, this criterion is not applicable.

- Assessment:**
- Met
  - Not met
  - Not applicable

**Criterion 28** *The official date of data release was announced before the release.*

Your assessment of the criterion goes here.

**Details Required:** Indicate whether the official date of data release for the annual release or releases of data to a significant stakeholder(s) was planned for and announced at least six months in advance.

- Assessment:**
- Met
  - Not met
  - Unknown
  - Not applicable

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

Your assessment of the criterion goes here.

**Details Required:** Indicate whether the data was released on or before the official date of data release. For data releases to a significant stakeholder(s), this criterion is to be based on the planned release date (as reported in the Calendar of Deliverables) versus the actual release date.

**Assessment:**       Met       Not met       Not applicable

**Criterion 30** *Data processing activities are regularly reviewed to improve timeliness.*

Your assessment of the criterion goes here.

**Details Required:** Provide information regarding the programs or systems that are used to prepare and analyze the data, as well as the schedule for ongoing review of these activities and their ability to produce timely data.

**Assessment:**       Met       Not met

## 2.2 Documentation Currency

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

Your assessment of the criterion goes here.

**Details Required:** Give information on the availability of data quality documentation for users once data or reports can be internally or externally accessed. Provide details on where and how this information can be accessed. The recommended components of good data quality documentation are further discussed in Section 4.

**Assessment:**       Met       Not met

**Criterion 32** *Major reports were released on schedule.*

Your assessment of the criterion goes here.

**Details Required:** Indicate whether the major reports for the data holding were released on schedule. Detail the circumstances around slippage of release dates, including steps taken to advise users.

**Assessment:**       Met       Not met       Not applicable

### 3. Comparability Dimension

#### 3.1 Data Dictionary Standards

**Criterion 33** *All data elements are evaluated to determine their inclusion within the CIHI Data Dictionary.*

Your assessment of the criterion goes here.

**Details Required:** Note results of review of all data elements in existing data holdings and data holdings being developed against the CIHI Data Dictionary. Indicate those that do not agree with the data dictionary standard, as well as data elements that do not currently have a standard in the CIHI Data Dictionary.

**Assessment:**       Met       Not met       Unknown

**Criterion 34** *Data elements from a data holding that are contained within the CIHI Data Dictionary must conform to dictionary standards.*

Your assessment of the criterion goes here.

**Details Required:** For data elements that are currently contained in the CIHI Data Dictionary, describe the level of conformance to data dictionary standards, as well as any justifiable deviations from CIHI standards. If the data dictionary does not contain any data elements found in the data holding, then the rating is not applicable.

**Assessment:**       Met       Not met       Unknown       Not applicable

#### 3.2 Standardization

**Criterion 35** *Data is collected at the finest level of detail practical.*

Your assessment of the criterion goes here.

**Details Required:** Indicate whether all core data elements are collected with the necessary detail required for publication and for linking or comparison purposes. Give explanations for any cases where data is not captured at the finest level of detail.

**Assessment:**       Met       Not met

**Criterion 36** *For any derived data element, the original data element remains accessible.*

Your assessment of the criterion goes here.

**Details Required:** Report on the presence and accessibility of original data elements in the database.

**Assessment:**       Met       Not met       Not applicable

### 3.3 Linkage

**Criterion 37** *Geographical data is collected using the Standard Geographical Classification (SGC).*

Your assessment of the criterion goes here.

**Details Required:** Report on the extent to which the entities on which data is collected (facilities, persons, province, etc.) are identifiable by either postal code (all six digits) or the relevant Standard Geographical Classification. If the lowest level of geography used is province, standard Canada Post province codes should be used. As geographical information can apply to more than one entity, clinical databases, for example, should collect geographical information not only on the patient, but the facility as well.

**Assessment:**  Met  Not met

**Criterion 38** *Data is collected using a consistent time frame, especially between and within jurisdictions.*

Your assessment of the criterion goes here.

**Details Required:** Indicate whether sufficient information between and within jurisdictions is available that would allow the data to be analyzed using a consistent time frame. Note any existing barriers or challenges to conducting such analyses.

**Assessment:**  Met  Not met

**Criterion 39** *Identifiers are used to differentiate facilities or organizations uniquely for historical linkage.*

Your assessment of the criterion goes here.

**Details Required:** Note the extent to which a unique code acceptable for historical linkage purposes (provincially assigned identifier or equivalent) exists for each facility or organization and is available on the database.

**Assessment:**  Met  Not met  Not applicable

**Criterion 40** *Identifiers are used to differentiate persons or machines uniquely for historical linkage.*

Your assessment of the criterion goes here.

**Details Required:** Note whether a unique person or machine identifier is available in the database that could be used to link to corresponding records across different time periods.

**Assessment:**  Met  Not met  Not applicable

### 3.4 Equivalency

**Criterion 41** *Methodology and limitations for crosswalks and/or conversions are documented.*

Your assessment of the criterion goes here.

**Details Required:** Describe any crosswalks and/or conversions employed by the data holding, indicating how often their methodology and limitations are documented for users. Provide the name of this documentation and where it may be accessed internally and externally.

**Assessment:**         Met         Not met         Not applicable

**Criterion 42** *The magnitude of issues related to crosswalks and conversions falls into one of the predetermined categories.*

Your assessment of the criterion goes here.

**Details Required:** Assess the efficacy of the crosswalks and conversions used in the database. Note testing procedures for new crosswalks or conversions and procedures to detect and deal with misclassifications. If the database uses more than one crosswalk or conversion, base the overall assessment on the weakest.

**Assessment:**         Minimal (no or few issues)  
                          Moderate (identifiable issues that are limited in scope)  
                          Significant (a significant portion of the data is not being converted properly and this has an impact on results)  
                          Unknown (equivalency has not been investigated)  
                          Not applicable

### 3.5 Historical Comparability

**Criterion 43** *Documentation on historical changes to the data holding exists and is easily accessible.*

Your assessment of the criterion goes here.

**Details Required:** Provide information regarding whether documentation of historical changes exists, is maintained in one document and how frequently it is updated. Discuss the content of the document, which should include changes to concepts, methodologies, frames and data elements.

**Assessment:**         Met         Not met

**Criterion 44** *Trend analysis is used to examine changes in core data elements over time.*

Your assessment of the criterion goes here.

**Details Required:** Indicate whether trend analysis was performed for core data elements since the last data quality assessment.

**Assessment:**       Met       Not met       Not applicable

**Criterion 45** *The magnitude of issues associated with comparing data over time falls into one of the predetermined categories.*

Your assessment of the criterion goes here.

**Details Required:** Discuss limitations or challenges involved in producing valid trend estimates.

**Assessment:**

- Minimal (no or few issues in producing comparable trends)
- Moderate (issues 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)
- Not applicable

## 4. Usability Dimension

### 4.1 Accessibility

**Criterion 46** *A final data set is made available per planned release.*

Your assessment of the criterion goes here.

**Details Required:** Indicate whether the data that is used for analysis and the creation of reports or releases is saved in a secure location for future reference. Discuss the format of the data and provide details on how and where it may be accessed.

**Assessment:**       Met       Not met

**Criterion 47** *Standard tables and analyses using standard format and content are produced per planned release or upon request.*

Your assessment of the criterion goes here.

**Details Required:** Provide information about commonly used standard tables and analyses, including how they are made available to users per planned release.

**Assessment:**       Met       Not met

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

Your assessment of the criterion goes here.

**Details Required:** Describe the CIHI dissemination systems used to define, catalogue and/or publicize products of the data holding per planned release.

**Assessment:**             Met             Not met

## 4.2 Documentation

**Criterion 49** *Current data quality documentation for users exists.*

Your assessment of the criterion goes here.

**Details Required:** Indicate whether a stand-alone data quality document for users (internal and external) is made available at least once a year. Note the format, location and update schedule for the document.

**Assessment:**             Met             Not met

**Criterion 50** *Current metadata documentation exists.*

Your assessment of the criterion goes here.

**Details Required:** Provide information regarding the status of metadata documentation for the data holding for internal purposes, as well as the practices for reviewing and updating the documentation. Please indicate where the internal metadata documentation may be accessed.

**Assessment:**             Met             Not met

**Criterion 51** *A caveat accompanies any preliminary release.*

Your assessment of the criterion goes here.

**Details Required:** Discuss any unofficial or official preliminary releases provided by the data holding for the time period. Include whether a caveat was provided and the substance of that caveat.

**Assessment:**             Met             Not met             Not applicable



## 4.3 Interpretability

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

Your assessment of the criterion goes here.

**Details Required:** Provide data holding practices that enable users (internal or external) to provide feedback to the program area, as well as mechanisms that facilitate contact from CIHI with major users.

**Assessment:**  Met  Not met

**Criterion 53** *Revision guidelines are available and applied per release.*

Your assessment of the criterion goes here.

**Details Required:** Indicate whether data holding–specific revision guidelines are available and applied whenever data is released or extracted. Note that the review schedule for these guidelines needs to be current and note where they may be accessed.

**Assessment:**  Met  Not met

## 5. Relevance Dimension

### 5.1 Adaptability

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

Your assessment of the criterion goes here.

**Details Required:** Discuss the liaison mechanisms data holding program area staff have in place to help stakeholders stay abreast of developments in the field.

**Assessment:**  Met  Not met

**Criterion 55** *The data holding is developed so that future system modifications can be made easily.*

Your assessment of the criterion goes here.

**Details Required:** Assess the ease of future modifications by reporting on whether the data holding has demonstrated the ability to adapt to an important emerging issue, to a new technical standard or to a major data quality limitation within the last year.

**Assessment:**  Met  Not met

## 5.2 Value

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

Your assessment of the criterion goes here.

**Details Required:** Provide information regarding the mandate of the data holding, the health care information gap it addresses and whether it has been recently assessed in relation to the other data holdings within CIHI and across the field externally.

**Assessment:**            Met            Not met

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

Your assessment of the criterion goes here.

**Details Required:** List the approaches taken by data holding area staff to monitor usage (for example, tracking the number and type of data requests, the use of eReports, high-profile uses of the data, web page hits, press clippings, news items, citations, staff-authored papers, sales, media appearances/contacts of staff, conferences and policy forums). Provide details on any particularly interesting usage seen or the most frequent type of data request.

**Assessment:**            Met            Not met

### **Criterion 58** *User satisfaction is periodically assessed.*

Your assessment of the criterion goes here.

**Details Required:** Describe the methods used to assess client satisfaction and their frequency of use. Note the general level of satisfaction reported by users, stakeholders and internal CIHI staff.

**Assessment:**            Met            Not met

## Appendix E—Subcategories for Metadata Documentation

As noted in Section 4.3, there are seven categories in CIHI’s metadata documentation for data holdings. Each of these categories is further defined by subcategories detailed below:

### Category: Database Description

Subcategory	Definition and/or Example
Background	Information for giving readers a general overview of the database (for example, national database for information on all acute care hospitals).
Purpose	Objective of the database (for example, developed to collect data on hospital discharges).
Scope	What are the subjects of inquiry and analysis that are of interest to users (for example, all provinces except Quebec)?
Stakeholders	Those who have a vested interest in the results of a database (for example, Ontario Ministry of Health and Long-Term Care).
Roles and responsibilities for different project team members, including data providers	Description (generic, if possible) of what each member does within the project team context (for example, data provider delivers most recent data to CIHI by a certain date in a certain format).
Definition of concepts	Description of database-specific terms used (for example, hospital separation = discharge or death of an inpatient).
List of internal contact persons	Those at CIHI who can be reached for further information on the database (contact coordinator, senior analyst, etc.).

### Category: Methodology

#### Methodology—Population of Interest/Reference

Subcategory	Definition and/or Example
Definition	Description of the population for which information is wanted/population that is available.
Exclusions	Description of records that are excluded from the population of interest for one reason or another.

## Methodology – Frame

<b>Subcategory</b>	<b>Definition and/or Example</b>
Definition	List of units that provide access to the population of reference (for example, facilities with acute inpatient beds in Canada other than Quebec).
List of data elements available	What data elements are available from the records on the frame?
List of units and contact persons	Which units are to be on the frame? The contact information of data providers when there are questions or concerns.
Historical changes to frame	Description of how the frame has evolved since its conception (for example, number of records, change in data providers, change in coverage).
Frame maintenance:	
– Adding/removing/merging units	Procedure in place to deal with units on the frame when there have been additions, deletions, amalgamations, etc.
– Updating units	Procedure in place to deal with data elements of records or records that have changed.

## Methodology – Survey Design (Survey Only)

<b>Subcategory</b>	<b>Definition and/or Example</b>
Sample design description/ explanation (sampling strategy, method used)	Description of why a sample was chosen, how it was chosen (that is, time/cost savings, randomly selected within each province, stratification of data elements).
Sampling units definition	Description of what is being considered, the sampling unit, for example, facility, chart (distinct and non-overlapping units into which the population is divided for the purpose of sample selection).
Sample selection:	
– Sample size calculation	Description of how the sample size was determined.
– Sample allocation	Description of how the sample was distributed among the strata (for example, 100 charts per province).
Probability of selection	Proportion of the number of sampled units/number of population units with each stratum.

## Category: Data Collection and Capture

Subcategory	Definition and/or Example
Data element list:	
– Accepted values	The validity edits that exist for each data element that is captured.
– Historical changes	Documentation detailing the data elements whose possible values have changed over time.
– Record layout	The expected record layout of the data file being received from the data provider.
– Changes to data elements	Documentation detailing the data elements that have changed over time (additions, deletions, becoming more detailed, less detailed).
– Data dictionary inventory	Contains data element names and definitions as well as the specifications associated with each data element.
Vendor or application specifications:	
– Record layout	Record layout required in order for data provider’s data file to be processed.
– Changes to specifications	Documentation of the changes made over the years in what the processing system requires from the data file.
– List of approved vendors (if applicable)	List of vendors currently used and any others that are recommended from past projects or being considered for future projects.
Client contact information:	
– Contact procedures	Procedures to contact clients to verify and update the information.
– Contact information	Listing of the required information to contact clients (name, title, phone number, email address).
– Data provider contact letter	Letter that provides contact information of the data provider.
– Log of contacts	Document that details all communication with clients (method of contact, who and when contacted, follow-up, etc.).
– Non-response follow-up	Procedure to follow when expected data (at record level or data-element level) is not present (letter, phone call, etc.).

<b>Subcategory</b>	<b>Definition and/or Example</b>
<b>Data submission specifications:</b>	
– Deadline	Expected date of data delivery by data provider.
– File naming convention	Expected convention for data file names coming from data provider.
– Detailed record layout	Expected record layout of files coming from data provider.
– Data collection method	How will data be delivered by data provider (online, mail, CD, diskette, etc.)?
<b>Database coding manual:</b>	
– Copy of abstract/form/questionnaire	Manual to include copy of questionnaires/forms used by the data provider to collect the information.
– Translation	Manual to include any translated documentation.
– Instruction to data providers	Manual to include instructions for the data provider to follow for data submission.
– List and description of acceptable values for mandatory/optional data elements	Manual to provide details on which data elements are required and what values are acceptable.
– List of coding standards (if applicable)	Manual to clearly specify what coding standards are expected to be followed by the data provider (for example, ICD-9, ICD-10-CA/CCI).
<b>Standardization of data elements derived from different data providers:</b>	
– Standardization procedures	Procedure to standardize values for data elements if not standard upon submission.
<b>Data consolidation into a common file:</b>	
– Process to “close” the database	Procedure to create common data file from several files received from data providers.

## Category: Data Processing

Subcategory	Definition and/or Example
Data flow diagram:	
– Programs	Documentation or diagram outlining the data processing programs.
– Input/output files	Documentation or diagram outlining the input files required for data processing programs and the output files resulting.
Editing:	
– List of valid codes for each data element	List of valid codes for data elements during each stage of data processing.
– Validity checks	Documentation of all the validity checks during each stage of data processing.
– Consistency checks	Documentation of all the consistency and distribution edit checks during each stage of data processing.
– Outlier detection checks	Documentation of all the outlier detection edits performed during each stage of data processing.
– Log of edit failures (file available)	File that contains the records or data elements that failed the various edits (record-identifying information, information describing error[s]).
– Minimal response data requirements	Criteria in place to ensure that quality is not jeopardized due to invalid or incomplete responses (out-of-scope, total non-response, partial response, complete response).
– Actions for flagged edit failures	Procedure for dealing with edit failures or missing data (contact data providers, imputation, reweighting).
– Editing process	Documentation of entire editing process.
Coding (open questions, closed questions with an “other” category)	Procedure for assigning codes for responses that are the result of open or closed questions.
Imputation:	
– Description of the imputation methods	If imputation is used, description of which methods are used to replace invalid or missing values.
– Fields	Which data elements can be/were imputed.
– Creation of imputed data elements	Description of how values are calculated using the imputation method to replace invalid or missing values.
– Log of imputations	File that contains the records or data elements that now contain imputed values and what method was used (if more than one can be used).
– Imputation process	Documentation of entire imputation process.

<b>Subcategory</b>	<b>Definition and/or Example</b>
Creation of derived and grouped data elements:	
– Description	Documentation of which data elements have been derived or grouped from others and for what purpose.
– Data element names	Documentation of the data element names that resulted from being derived or grouped.
– Source	Documentation of any data element that was used as a basis for the derivation or grouping.
– Calculation	Detailed documentation in transforming the original data into a derived or grouped data element.
– Log of historical changes	File that contains all the derivation and groupings of data elements that have occurred over the years.
– Process for creating derived data elements	Documentation of the process to derive or group data elements.
– Process to change the current derivation process	Documentation for the process to follow if changes in derivation or grouping are to occur.

---

## Data Processing—Weighting (Survey Only)

<b>Subcategory</b>	<b>Definition and/or Example</b>
Weighting methodology:	
– Sampling weight	List of sampling weights for each sampled record (number of units in the population each sampled unit represents).
– Weight adjustments	Process to adjust weights to account for non-response, to improve estimates, etc.
– Final weight	List of final weights for each sampled record (the combination of sampling weight and weight adjustments).
Weighting process	Documentation about how all weights are calculated.



## Data Processing—Production of the Analytical File

<b>Subcategory</b>	<b>Definition and/or Example</b>
Creation of the file(s) (or tabulation file in the case of a survey)	Documentation of steps to follow for producing the file that will be used to compute estimates (steps to follow, structure and linkage, file format).
Location	Documentation specifying the location where the file that will be used to compute estimates can be accessed.
Detailed record layout(s) (collected and derived data elements)	Documentation specifying the record layout of the file that will be used to compute estimates.
Sample letter and documents (CD/diskette/flat file) sent to data providers for review and validation	Letter(s) or document(s) sent to data providers to review the estimates that have been obtained from an analysis.
Data access (who, how, what kind of access?)	Protocol exists for accessing the resulting estimates.
Description/explanation of the estimation method and variance estimation method used (survey only)	Documentation of methodology to calculate estimates and variances.
Process for creating estimates	Documentation for calculation of estimates or results.

## Category: Data Analysis and Dissemination

<b>Subcategory</b>	<b>Definition and/or Example</b>
Table production:	
– List of commonly produced tables	Documentation of the various output tables produced (for publication, etc.).
– Table templates	Document contains templates for the various tables that are produced.
– Historical tables	Document containing the tables that have been produced over the years (for comparative purposes, ad-hoc requests, etc.).
– Programs to produce tables	Document indicating the location of programs and their purpose for table creation.
– Process for table creation	Up-to-date document that describes the entire table creation process.
– Process for table validation	Document that lists the various validation checks to ensure that the tables are populated with acceptable responses.

<b>Subcategory</b>	<b>Definition and/or Example</b>
<b>Suppression of confidential information:</b>	
– Methods of suppression used	Description of the method of suppression used (cell suppression, recoding into intervals, top/bottom coding, rounding).
– Suppression criteria	Description of the criteria that are not met for a cell to be suppressed (for example, based on too few responses).
– Encryption	Description of the encryption method used (for example, for health card numbers).
– Process for suppression	Documentation of the entire cell suppression process.
<b>Report creation:</b>	
– Template for commonly produced reports	Document contains templates for the various reports produced.
– Previously produced reports	Document containing the reports that have been produced over the years (for comparative purposes, etc.).
– Process for creating reports	Documentation of the entire report creation process.
– Process for validating reports	Document that lists the various validation checks in order to ensure that the resulting reports contain the correct content.
– Sign-off by data providers	Documentation to ensure data providers have signed off on all reports and the process for signing off.
– Data release process	Documentation of the data-release process.
<b>Publications:</b>	
– List and description of available products and services	Detailed list of what services and publications are available for users.
– Methodological notes	Inclusion of a methodological notes section in the publication (including a section on data quality).
– Sample data announcement (including contact information)	Publication mentions when preliminary estimates have been released.
– Frequently asked questions	Publication includes an FAQ section to quickly aid users.
– Process for publication	Documentation for the entire publication process (who to contact, etc.).

<b>Subcategory</b>	<b>Definition and/or Example</b>
Ad-hoc queries:	
– Data request form	A data request form exists for users for future requests.
– List of queries done	A list exists of which requests have been answered.
– Saved queries	A list exists of which requests have not yet been answered and are still being used to respond to queries.
– Program(s) used to generate the results	Documentation of the programs used in order to respond to data requests.
– Process for data requests	Documentation of the entire process for responding to data requests (who is responsible, deadline, etc.).
– Process for validation of queries	Documentation of validating the output from the data request prior to making it available to clients (independent verification).

## Category: Data Storage

<b>Subcategory</b>	<b>Definition and/or Example</b>
Data storage requirements	Documentation of the requirements for storing data.
Historical files (location and record layout)	Documentation of where and what as it pertains to historical files.
Data dictionary inventory	Contains data element names and definitions as well as the specifications associated with each data element.

## Category: Documentation

<b>Subcategory</b>	<b>Definition and/or Example</b>
Chronological notes:	
– Historical data limitations	Documentation of the limitations of historical data.
– Comparability with other sources	Documentation concerning the comparability with other sources.
– Changes in methodology/ collection/processing	Documentation outlining any changes in methodology, collection and processing over the years.
Documentation specific to different users with different level of detail	Different types of documentation exist for different types of users (analysts, researchers). Includes data quality reabstraction or special study reports as well as data quality assessment reports for such studies.

<b>Subcategory</b>	<b>Definition and/or Example</b>
Process to create documentation for different users	Process exists to produce specific documentation for specific users (who is responsible, when it is to be done, validating the report, data quality contact person, etc.).
Data quality assessment report	Data quality assessment report for latest reporting period has been produced.
Process to create data quality assessment report	Process has been set up to complete the data quality assessment report (who is responsible, when it is to be done, validating the report, data quality contact person, etc.).
Data quality documentation for external users	Data quality documentation for external users is produced. Includes data quality reabstraction/special study reports as well as data quality assessment reports for such studies.
Process to create data quality documentation for external users	Process has been set up to complete the data quality documentation for external users (who does it, what is the deadline, validation, etc.).
Education sessions/workshops	Education sessions or workshops are organized for external clients and/or internal staff to educate and answer questions concerning the data holding.
Gantt chart—steps, start/end dates, person responsible	Tool used to assist in making sure all required activities are on schedule, completed and who is involved.
Process to create/update Gantt chart	Process exists to create/update the Gantt chart so that relevancy is maintained (who is responsible, how frequent, communicate to project team).

## Appendix F—Glossary

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

**accessibility:** A characteristic. It is defined as the ease with which a data holding's data can be obtained from CIHI.

**accuracy:** A dimension. How well information in the data holding, or derived from the holding, reflects the reality that it was designed to measure.

**adaptability:** A characteristic. The degree to which a data holding 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, and statistics used to summarize (or to aggregate) the data, such as the mean or median.

**annual release:** The standard set of data that is provided on a yearly basis to those that use the data. This can include, for a particular reference period, an entire database or a subset of the database.

**assessment tool:** The core component of the data quality framework. It comprises 61 different criteria that are used to identify aspects of concern with relation to data quality and to assess the limitations and strengths of a data holding.

**bias:** An assessment to what degree systematic differences occur between reported values in a data holding 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*.

**CCI:** Canadian Classification of Health Interventions.

**characteristic:** An aspect of data quality that comprises one criterion or more.

**CIHI Data Dictionary:** The data quality 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 make 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 data holding 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. It is defined as the degree to which the frame of a database describes the population of reference.

**criterion:** A specific statement that relates to a detailed element of data quality. Each criterion is given a rating of *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. 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 data holding.

**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 relates.

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

**data flow:** The path data takes as it is processed. It is 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 data holding. Data providers normally perform the function of data capture.

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

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

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

**de-identified:** Refers to data which has information removed so that users will not be able to identify specific individuals, institutions or medical equipment from it.

**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. It is defined as information on a data holding's 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 reports for the data holding 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 data holding.

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

**editing:** The process that checks for and identifies missing, incorrect or invalid data.

**equivalency:** A 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 provide advice 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 treating similar items by placing them into a category designed specifically for them.



**historical comparability:** A characteristic. It is defined as the consistency of data concepts and methods over time that allow one to make valid comparisons from different time periods.

**health care provider:** An individual who has delivered, who is delivering or who has the potential to deliver health care–related services or goods. At the time of writing for this version of the CIHI Data Quality Framework, it was the only finalized concept within the CIHI Data Dictionary.

**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 Statistical Classification of Diseases and Related Health Problems, 9th Revision. It is defined as a set of internationally accepted codes for classification of medical diagnoses and conditions.

**ICD-10-CA:** Canadian enhanced version of the World Health Organization’s ICD-10. It is labelled as the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Canada.

**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.

**inter-rater reliability:** Discrepancy rate that is calculated when two expert reabstractors code the same item to determine if the responses are consistent between them.

**intra-rater reliability:** Discrepancy rate when one expert reabstractor must code the same item twice to determine the consistency of this particular reabstractor in recording information.

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

**item non-response:** A characteristic. It is also referred to as partial non-response. It includes blank data elements found on a record received that should not be blank.

**item non-response rate:** The percentage of data elements for which data was 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 pairwise agreement among a set of coders that are making category judgments and corrects for expected chance agreement. The formula is the following:

$k = (\text{observed count of agreement} - \text{expected count of agreement}) / (\text{total number of respondent pairs} - \text{expected count of agreement})$ . The SAS procedure (PROC FREQ with the /AGREE option) also calculates the kappa statistic.

**linkage:** A characteristic. How easily the data holding can be linked to other holdings.

**linking data element:** Refers to a data element that is common to two or more data sets and may be used to combine them.

**longitudinal data:** Data that spans 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 data holding and the procedures that are used to collect and process the data.

**measurement error:** A characteristic. It is defined as the difference between the value that is reported in the data holding and the true but unknown value that should have been reported.

**met:** An assessment rating. It 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 to run a database.

**microdata:** The data that is used for analysis and the creation of reports.

**minimal:** An assessment rating. It signifies that the level being measured is a low one.

**missing at random:** Data that is absent from a data holding on a random basis (the fact that it is missing is not related to any other data element).

**moderate:** An assessment rating. It signifies that the level being measured is a medium one.

**none:** An assessment rating. It signifies that there are no occurrences of the attribute being measured.

**non-response:** Failure to obtain data on all the units in 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:** An assessment rating. It signifies that the requirements for the criteria cannot be achieved.

**not met:** An assessment rating. It 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 is 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 data holding. 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 data holding 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 in which units that are not part of the population of reference are included in 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:** It consists of either the application of programs or a sequence of procedures to finalize a database. Processing can be done for many reasons, including producing estimates, totals or frequencies or for error testing.

**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.

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

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

**rate of under-coverage:** A rate calculated from the number of units not in 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 arrives from data providers and is not modified.

**reabstraction study:** A study in which designated experts go through the same data-collection process normally done by the data providers. 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 relates and the end of the reference period is the last date to which the data relates.

**release:** Any report, data release or output from the data holding.

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

**reliability:** See *consistency*.

**response burden:** A measure of how difficult it is for data providers 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 has 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).

**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. Standard data elements or consistent reporting periods are examples.

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

**Standard Geographical Classification (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. It is defined as 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 is at the time of release. The currency of the data is measured in terms of the gap between the end of the reference period to which the data pertains and the date on which the data becomes available to users.

**under-coverage:** A situation in which 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 in the frame.

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

**unknown:** An assessment 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 data holding's data may be interpreted, understood and accessed. The issue of usability also relates to potential users knowing of a holding's existence and the data being in a readily accessible, user-friendly form.

**usage:** The extent to which the data from a data holding is 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:** A characteristic. It is defined as the contribution of a data holding to population 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.

## Bibliography

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

Statistics Canada. *Policy on Informing Users of Data Quality and Methodology (Approved March 31, 2000)*. From <[http://www.statcan.gc.ca/about-apercu/policy-politique/info\\_user-usager-eng.htm](http://www.statcan.gc.ca/about-apercu/policy-politique/info_user-usager-eng.htm)>.

Statistics Canada. *Statistics Canada Quality Guidelines* (fifth edition). Ottawa, Ont.: Statistics Canada, 2009. Catalogue no. 12-539-XWE.

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

U.S. Census Bureau. *Definition of Data Quality* (version 1.3). (June 2006.) From <[http://www.census.gov/quality/P01-0\\_v1.3\\_Definition\\_of\\_Quality.pdf](http://www.census.gov/quality/P01-0_v1.3_Definition_of_Quality.pdf)>.

