# A New Method for Database Data Quality Evaluation at the Canadian Institute for Health Information (CIHI)

(Practice-Oriented Paper)

**Jennifer A. Long**
Consultant, Data Quality
Canadian Institute for Health Information
Email: jlong@cihi.ca
www.cihi.ca

**Craig E. Seko**
Senior Analyst, Health and Outcomes Statistics
Health Statistics Division, Statistics Canada
Email: sekocet@statcan.ca
www.statcan.ca

**Abstract**: Information quality (IQ) problems can have severe consequences in the health care sector. Since its inception, the Canadian Institute for Health Information (CIHI) has recognized the importance of information quality and has implemented a new method designed to evaluate the data quality of the numerous CIHI data holdings. The goal of evaluation is to identify data quality priorities for the purpose of data quality improvement. To date, six database evaluations have been conducted and it appears that the evaluation process has been successful in meeting its primary objective. It is concluded that the new method is a useful tool for data quality improvement, especially in the health care sector where data quality improvement can result in better health information and, ultimately, better health.

**Key Words**: Data Quality, Evaluation, Database Evaluation, Data Quality Framework

# Introduction

Information quality (IQ) problems can have severe financial and operational consequences for organizations [13]. In the health care sector, faulty hospital patient information can lead to incorrect diagnoses and interventions, with potentially fatal consequences. Since its inception, the Canadian Institute for Health Information (CIHI) has recognized the importance of data quality. Although data quality has been a priority since the establishment of CIHI, a new method has been recently implemented to systematically evaluate and improve the data quality of the CIHI data holdings. Some preliminary results based on the method are now available. This paper outlines the data quality evaluation methodology, provides a summary of the available data that result from the process (i.e., data quality meta-data), and offers some early insights into the process.

# Background

Incorporated in 1993, CIHI is a federally chartered, independent, not-for-profit organization. CIHI has taken a central role in the development of Canada's health information system. CIHI's mandate is to "serve as the national mechanism to coordinate the development and maintenance of a comprehensive and integrated health information system for Canada" and "to provide and co-ordinate the provision of accurate and timely information required for: the establishment of sound health policy, the effective management of the Canadian Health System, and for generating public awareness about factors affecting good health" [3].

CIHI was established through the amalgamation of two non-governmental organizations, i.e., the Hospital Medical Records Institute (HMRI) and the MIS Group, along with selected databases and functions from the Health Information Division of Health Canada and the Health Statistics Division of Statistics Canada (STC) [9, 15]. In addition to the numerous data holdings inherited at the time of incorporation, and consistent with its mandate, an important role taken on at CIHI is the collection, processing, and maintenance of a growing number of administrative clinical databases or patient registries, as well as health human resources, health services, and health expenditures databases. To date, the CIHI data holdings include 22 administrative databases and patient registries, many of which are national in scope [10].

The challenge of overseeing over 20 data holdings is complicated by the fact that database methods are not standardized and data quality varies across the holdings. Contributing to the gravity of the challenge is the fact that CIHI data are used to allocate health care funding and resources as well as for comparative reporting. In recognition of the vital importance of data quality, as well as due to the responsibility of maintaining 22 databases or registries, the need for a standard systematic strategy to identify and solve data quality problems and to enable Senior Management to allocate finite resources for database improvement, has fast become imperative.

In response to the identified need for a standard, organized, and systematic approach to administrative data quality, one of the authors (Seko) was seconded from STC in 1999 to develop a data quality strategy in collaboration with CIHI Senior Management. The resulting strategy has a two-pronged approach. The first prong is the CIHI Data Quality Framework (Version 1) (CIHI-DQF (v1)) and the second prong is a large three-year special study of the most used CIHI administrative database (Hospital Discharge Abstract Database). This database contains selected personal and health information from every in-patient discharge everywhere in Canada except Québec and parts of Manitoba. Ambulatory care in some facilities is also included in some provinces. The special study, which started September 2000, is still ongoing and the year one results are summarized elsewhere [11].

The CIHI Data Quality Framework (Version 1) (CIHI-DQF (v1)), on the other hand, is the cornerstone of the new CIHI data quality strategy. The CIHI-DQF (v1) draws on the statistical literature [5, 4, 1, 2, 14] the STC Quality Guidelines [12], the Information Quality literature [7], the CIHI mandate, as well as the principle of Continuous Quality Improvement (CQI). The objectives of the CIHI-DQF (v1) are:

1) to standardize information on data quality, both for internal and external users;
2) to provide a common strategy for assessing data quality; and
3) to define a work process for CIHI's data holdings that identifies data quality priorities and produces continuous improvement in data quality [3].

To achieve these objectives, the framework outlines a standard multi-step "Quality Cycle" for the Institute as well as clear roles, responsibilities, and deadlines for the implementation of the cycle.

The Quality Cycle entails:

1) deciding on time period for the cycle, e.g., annually, biannually, etc…;
2) deciding on a date of data availability and publishing the date;
3) allocating resources for ***data quality analysis***, ***data quality evaluation***, and ***data quality documentation*** within the specified timeframe;
4) allocating time to investigate improvements;
5) adjusting operational plans if necessary; and
6) returning to step 2 once the cycle is completed [3].

In order to execute the Quality Cycle, support and several tools are available for each step that were described and assessed in a previous study [8].

Following ***data quality analysis*** in the Quality Cycle is ***data quality evaluation***, which is the focus of this paper. While the objectives of the CIHI-DQF (v1) are essentially to standardize efforts and to define a work process, the objectives of the database evaluation process are: 1) to identify and rank aspects of data quality needing improvement; and 2) to produce information on data quality that feeds into the creation of data quality documentation for users [3]. In other words, the purpose of database evaluation is to identify areas in need of improvement and to facilitate the documentation of limitations for users. Whereas the evaluations are internal documents since they may contain confidential information, any limitation detected via data quality analysis or evaluation is made available for users in the form of a standardized stand-alone ***data quality document*** [3].

According to the CIHI-DQF (v1), the concept of data quality is operationalized as a multi-level model and the framework includes an instrument and scoring algorithm for data quality measurement. The framework was ready and implementation started in April 2000. The deadline to implement the framework's Quality Cycle into operational plans was July 1, 2001 and the deadline for one evaluation per database is July 1, 2002. To date, seven major database evaluations have been conducted, two of which have been conducted on the same database. Although some of the evaluations have not yet been fully completed, they have been included for the purposes of this study.

# Rationale and purpose

While other statistical or health database institutes have standard data quality strategies in place [2], CIHI may be unique in using a measurable definition of data quality for the purpose of standard database assessment and improvement. Comparable results between databases enable operational planning for continual overall quality improvement. Although the evaluations are confidential and the results to date are preliminary, it was decided that the results could be shared for the purpose of contributing to the field of information quality as long as the individual databases were not identified. Hence, the purpose of this study was: to outline the CIHI database evaluation process; to present a simple descriptive summary of some results based on the process; and to offer some early insights into the process.

# Methods

At the time of this study, seven data quality evaluations were available. As two of the seven were conducted on the same database it was decided that only the most recent evaluation would be included, therefore six evaluations were made available for study. Five of the six evaluations were conducted and submitted during 2001 or early 2002 and one was submitted during August 1999. Two of the six evaluations were conducted by the authors, one was conducted by a Senior Analyst from the CIHI Data Quality Section, one was conducted by an independent consultant, and the remaining two were conducted

by staff in the corresponding database areas. Although the Data Quality Section independently validates[1] each database evaluation, at the time of the study three evaluations had not yet been validated. Each evaluation was conducted on a subset of data and the date spans of each subset were not necessarily synchronized. Only three of the six database subsets spanned the same date ranges. Specifically, the data evaluated from three of the six subsets spanned April 1, 2000 to March 31, 2001 (i.e., fiscal year 2000/2001) and the remaining three subsets spanned fiscal years 1999/2000 and 2000/2001, 1998/1999 (fiscal year), and 1998 (calendar year), respectively.

Other than noting that the databases represent different sectors of the health care system and some are national in scope, virtually no other information that can be used to identify the databases is provided because the evaluations are considered confidential and for internal use only. Accordingly, an arbitrarily assigned number from one to six identifies the databases involved. Again, note that any known data quality limitations are described separately for users.

## *Data quality*

According to the CIHI Data Quality Framework (Version 1) (CIHI-DQF (v1)), 'quality' is defined as 'fitness for use' and 'data quality' is operationally defined and measured along several commonly accepted and widely used data quality concepts and characteristics. Specifically, the concepts of database accuracy, timeliness, comparability, usability, and relevance are used to define overall data quality. Each of these concepts is, in turn, defined by many characteristics that are typically associated with them.

In particular, the characteristics of over-coverage, under-coverage, simple response variance, reliability, correlated response variance, collection and capture, unit non-response, item non-response, edit and imputation, processing, and estimation are used to define the concept of accuracy. Timeliness is defined by the difference between planned and actual release dates and the concept of comparability is defined by the comprehensiveness, integration readiness, standardization, equivalency, linkage-ability, product comparability, and historical comparability of the data. Usability is defined by the state of data accessibility, documentation, and interpretability, and relevance is defined by how well the data can be adapted to user's needs, and how valuable to users the data are. The concepts and characteristics are organized into a model that is used to operationalize data quality for the purpose of measurement and ultimately for database improvement. More detailed concept and characteristic definitions are provided in the CIHI Data Quality Manual, April 2001.

## *Data quality measurement*

More specifically, the CIHI Data Quality Framework (Version 1) (CIHI-DQF (v1)) operationalizes data quality as a four-level conceptual model. At the foundation of the model are 86 basic unit items that are known as criteria. The 86 criteria can be rolled-up using the framework algorithm into the second level of 24 characteristics (e.g., under-coverage, reliability, and interpretability) that in turn, can be rolled-up using the algorithm into the five dimensions of data quality (i.e., accuracy, timeliness, comparability, usability, and relevance). Finally, the five dimensions can be reduced using the algorithm into one overall database evaluation. Figure 1 below provides a summary of the four-level conceptual model.

[1] See Methods - *Data quality measurement* for more information on evaluation validation

Overall Evaluation*
↑
5 Data Quality Dimensions*
1. Accuracy 2. Timeliness 3. Comparability 4. Usability 5. Relevance
↑
24 Data Quality Characteristics*
↑
86 Data Quality Criteria**
↑
Data Quality Analysis and Database Document Compilation

*0. not applicable, 1. unknown, 2. not acceptable, 3. marginal, and 4. appropriate
**0. not applicable, 1. unknown, 2. not met, and 3. met

**Figure 1. The CIHI Data Quality Framework (Version 1) Four-Level Model**

Database evaluations are submitted as evaluation reports to the Data Quality Section for validation. As part of an evaluation report, each criterion must be addressed in writing and all scores must be substantiated in writing. As part of the validation process, the Data Quality Section ensures that the suggested data quality analyses have been run, that each criterion has been addressed, and that criterion interpretation and scoring is as standard as possible. The Data Quality Section also makes certain that recommendations based on the evaluation process are included and prioritized. Typically, revisions are requested from, and the final scores are decided in collaboration with, the Data Quality Section.

Upon report completion, the criterion scores are transcribed onto the CIHI Data Quality Evaluation Instrument (Version 1) (CIHI-DQF Instrument (v1)), which is essentially the four-level model reformatted as a questionnaire. The purpose of the CIHI-DQF Instrument (v1) is to facilitate the computation and tracking of database evaluation scores. The 86 criteria (or level 1) appear on the CIHI-DQF Instrument (v1) as checklist type items that are scored using a four-point ordinal scale as either "not applicable" (0), "unknown" (1), "not met" (2), or "met" (3). Figure 2 is an excerpt of the instrument that illustrates how the characteristic of documentation appears as well as the constituent criteria for documentation.

| | | |
|---|---|---|
| 4.2 Documentation (i.e., existence, completeness, etc. of data quality documentation for users (internal documentation is part of processing.)) Score: | 4.2.1 Data quality documentation (documents) exists per release (reference) period | 0.Not Applicable / 1.Unknown / 2.Not Met / 3.Met |
| | 4.2.2 Documents are in standard form (to be revised) per release (reference) period<br>• collection forms are included<br>• interaction with other CIHI databases and registries is explained<br>• mandate and purpose of the database/registry is given<br>• a description of procedures to reduce errors is given<br>• descriptions of concepts, definitions, and methods are given; universe, population, and frame are explained<br>• a description of any changes from historical procedures is given<br>• data quality issues (strengths and limitations) that might affect data use are highlighted.<br>• changes known for future releases are given | 0.Not Applicable / 1.Unknown / 2.Not Met / 3.Met |
| | 4.2.3 Documents are explicitly referenced in all output per release (reference) period | 0.Not Applicable / 1.Unknown / 2.Not Met / 3.Met |

**Figure 2. An excerpt from The CIHI Data Quality Framework Evaluation Instrument (Version 1)**

Once the criteria have been scored, the algorithm is used to compute the scores for the 24 characteristics and five dimensions, as well as for the database overall. Unlike the checklist type criteria, the characteristic, dimension, and overall scales are scored as either 0 = "not applicable", 1 = "unknown", 2 = "not acceptable", 3 = "marginal", or 4 = "appropriate" (i.e., five-point ordinal scales). The instrument is organized by dimension and within each dimension are the constituent characteristics and corresponding criteria. For instance, the first forty-one criteria provide the base for the first eleven characteristics that in turn are the foundation for the concept of database accuracy.

Either the evaluator or the Data Quality Section transcribes the criteria. Once transcribed onto the instrument, the data are entered and checked for missing data or errors. Once checked and, if necessary, corrected, the data are read into a statistical data set and the characteristic, dimension, and overall scores are computed. The algorithm used to generate the scores is published in the internal CIHI Data Quality Manual, April 2001.

The measurement properties of the instrument are unknown; however, as previously stated the Data Quality Section independently validates all evaluation scores. To date three of the six evaluations have been checked and the remaining three are scheduled for validation. It should also be noted that results were not available at the criterion level for three of the evaluations and due to this lack of data the level 1 or criteria scores were not summarized. For these cases, only characteristic or higher-level scores were provided. It is anticipated that substantiated criteria scores will be available for all future evaluations.

While the algorithm serves as an excellent guide, it must be emphasized that all final scores are decided upon in collaboration with the database area and the Data Quality Section. For the six evaluations included for study, in some instances scores could have been overridden. Overrides are permitted as long as they are based on informed, defensible reasoning and are documented. All overrides are conducted using code and no raw data are changed. In the interest of keeping the data quality measurement process as standard as possible it is obviously important to not use overrides unless absolutely necessary. It is anticipated that as the framework and algorithm develop, fewer and fewer overrides will be necessary.

## *Analysis*

A simple descriptive analysis was conducted for the evaluation scale scores. In addition to the results being presented for the individual databases, despite representing different timeframes the results were also combined in order to showcase how data quality might be tracked for the Institute as a whole. Specifically, the five-point ordinal scale scores (i.e., 0 (not applicable), 1 (unknown), 2 (not acceptable), 3 (marginal), or 4 (appropriate)) are presented for each database individually and are summarized across the holdings. Although the total number of observations is small, percentages are used to describe the data because the number of observations will soon be updated and, once updated, percentages will be used for comparisons across time.

Although a debatable practice, the ordinal ratings were also summarized as continuous data. In other words, the ordinal scale scores were combined to yield average data quality scores. It is recognized that this approach is open to question because the conceptual distance between the small number of ordinal categories is not uniform or is tenuous at best (e.g., the difference between unknown = 1 and not acceptable = 2). On the other hand, the average ordinal scales are arguably useful because they provide a quick summary. Where the ordinal scores were summarized as continuous data, the category of 0 = "not applicable" was removed and a four-point ordinal scale (i.e., 1 (unknown), 2 (not acceptable), 3

(marginal), or 4 (appropriate)) was used instead. In other words, all observations with scores equal to "not applicable" were removed.

What would qualify as a significant difference in data quality within databases across time or between databases has not been defined. Currently, any difference in scores must be validated and is considered important. When more data become available summary statistics will be calculated for the 86 criteria and more informative graphical summaries will be used, e.g., side-by-side box plots. It should also be noted that evaluation scores might change after independent verification so future summaries of the same evaluations may change. All data manipulation and analysis was conducted using the SAS System for Windows (v. 8.02).

# Results

Six data quality database evaluations, based on the CIHI Data Quality Framework, Version 1 (CIHI-DQF (v1)), were summarized for the purpose of this study. The database evaluation results, expressed as ordinal scale scores ranging from 0 (not applicable) to 4 (appropriate), are displayed in Table 1. Although the results are considered preliminary and not validated, at the time of this study almost all the data quality characteristics for database 1 were scored as "appropriate" except simple response variance, correlated response variance, and timeliness. Of the five dimensions, only timeliness was scored less than appropriate and the data quality of database 1 overall was found to be "appropriate".

Whereas the overall data quality of databases 4 and 6 was also found to be "appropriate", database 3 was rated as "marginal" overall, and databases 2 and 5 were evaluated as "not acceptable". In general, lower characteristic scores tended to result in lower dimension and overall scores.

| Database | Characteristics* | | | | | | | | | | | | | | | | | | | | | | | Dimensions* | | | | | Overall Score* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Over-coverage | Under-coverage | Simple response variance | Reliability | Correlated response variance | Collections and capture | Unit non-response | Item non-response | Edit and imputation | Processing | Estimation | Timeliness | Comprehensiveness | Integration | Standardization | Equivalency | Linkage-ability | Product comparability | Historical comparability | Accessibility | Documentation | Interpretability | Adaptability | Value | Accuracy | Timeliness | Comparability | Usability | Relevance | |
| 1 | 4 | 4 | 1 | 4 | 1 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 4 | 4 |
| 2 | 3 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 0 | 4 | 4 | 4 | 4 | 1 | 3 | 0 | 0 | 1 | 3 | 2 | 4 | 4 | 2 | 4 | 3 | 2 | 4 | 2 |
| 3 | 3 | 4 | 3 | 3 | 1 | 4 | 1 | 4 | 4 | 4 | 0 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 3 | 3 | 4 | 4 | 3 | 4 | 3 | 3 | 4 | 3 |
| 4 | 4 | 4 | 1 | 1 | 4 | 4 | 4 | 4 | 4 | 4 | 0 | 4 | 3 | 4 | 4 | 4 | 3 | 4 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 4 |
| 5 | 2 | 3 | 1 | 4 | 3 | 3 | 1 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 3 | 3 | 3 | 4 | 4 | 2 | 3 | 3 | 3 | 4 | 2 |
| 6 | 4 | 3 | 1 | 1 | 4 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 1 | 1 | 3 | 3 | 3 | 1 | 1 | 4 | 4 | 1 | 3 | 1 | 4 |

*0 = not applicable, 1 = unknown, 2 = not acceptable, 3 = marginal, and 4 = appropriate

**Table 1. Data quality characteristic, dimension, and overall scores of six data holdings**

As shown in Tables 2 and 3, the evaluation results, despite representing different time spans, were summarized by characteristic and by dimension, respectively, for CIHI as a whole. In terms of coverage,

three of the six databases received an "appropriate" rating for over-coverage and three of the six databases received an "appropriate" rating for under-coverage. Similarly, collection and capture, unit non-response, item non-response, edit and imputation, and processing tended to be scored as "appropriate".

In addition, most of the databases (4 out of 6) were rated as demonstrating "appropriate" timeliness and the characteristics of comprehensiveness, integration, standardization, equivalency, product comparability, accessibility, and adaptability also tended to be rated as "appropriate". However, the characteristics of simple response variance, reliability, correlated response variance, linkage-ability, historical comparability, documentation, and interpretability tended to receive less positive scores. Estimation was found to be "not applicable" for 3 of the 6 databases.

For the characteristics altogether, the modal score was "appropriate" (85/144 (59%)). Second to the most frequent score was the score of "marginal" (28/144 (19%)), followed by "unknown" (21/144 (15%)) and 10 of the 144 characteristic scores were categorized as either "not applicable" or "not acceptable" (Table 2).

| Characteristic | Not applicable (0) | Unknown (1) | Not acceptable (2) | Marginal (3) | Appropriate (4) | Total number of scores |
|---|---|---|---|---|---|---|
| Over-coverage | 0 (0.0%) | 0 (0.0%) | 1 (16.7%) | 2 (33.3%) | 3 (50.0%) | 6 |
| Under-coverage | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 3 (50.0%) | 3 (50.0%) | 6 |
| Simple response variance | 0 (0.0%) | 5 (83.3%) | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 6 |
| Reliability | 0 (0.0%) | 3 (50.0%) | 0 (0.0%) | 1 (16.7%) | 2 (33.3%) | 6 |
| Correlated response var. | 0 (0.0%) | 3 (50.0%) | 0 (0.0%) | 1 (16.7%) | 2 (33.3%) | 6 |
| Collection and capture | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 2 (33.3%) | 3 (50.0%) | 6 |
| Unit non-response | 0 (0.0%) | 2 (33.3%) | 1 (16.7%) | 0 (0.0%) | 3 (50.0%) | 6 |
| Item non-response | 0 (0.0%) | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 5 (83.3%) | 6 |
| Edit and imputation | 0 (0.0%) | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 5 (83.3%) | 6 |
| Processing | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 0 (0.0%) | 5 (83.3%) | 6 |
| Estimation | 3 (50.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 3 (50.0%) | 6 |
| Timeliness | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 2 (33.3%) | 4 (66.7%) | 6 |
| Comprehensiveness | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 1 (16.7%) | 5 (83.3%) | 6 |
| Integration | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 6 (100.0%) | 6 |
| Standardization | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 6 (100.0%) | 6 |
| Equivalency | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 0 (0.0%) | 5 (83.3%) | 6 |
| Linkage-ability | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 4 (66.7%) | 2 (33.3%) | 6 |
| Product comparability | 1 (16.7%) | 1 (16.7%) | 0 (0.0%) | 0 (0.0%) | 4 (66.7%) | 6 |
| Historical comparability | 1 (16.7%) | 1 (16.7%) | 0 (0.0%) | 2 (33.3%) | 2 (33.3%) | 6 |
| Accessibility | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 2 (33.3%) | 3 (50.0%) | 6 |
| Documentation | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 4 (66.7%) | 2 (33.3%) | 6 |
| Interpretability | 0 (0.0%) | 0 (0.0%) | 1 (16.7%) | 3 (50.0%) | 2 (33.3%) | 6 |
| Adaptability | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 0 (0.0%) | 5 (83.3%) | 6 |
| Value | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 0 (0.0%) | 5 (83.3%) | 6 |
| Totals | 5(3.5%) | 21(14.6%) | 5(3.5%) | 28(19.4%) | 85(59.0%) | 144(100.0%) |

**Table 2. Data quality characteristic scores**

Table 3 illustrates the six database evaluation scores rolled-up to the dimension level for the Institute as a whole. Whereas relevance and timeliness were, for the most part, found to be "appropriate", accuracy was found to be "appropriate" in three of the six databases and usability and comparability were often rated as "marginal". In total, for the five data quality dimensions the modal score was "appropriate" (15/30 (50%)) followed by "marginal" (10/30 (33%)), "not acceptable" (3/30 (10%)), and "unknown" (2/30 (7%)) (Table 3).

| Dimension | Not applicable (0) | Unknown (1) | Not acceptable (2) | Marginal (3) | Appropriate (4) | Total number of scores |
|---|---|---|---|---|---|---|
| Accuracy | 0 (0.0%) | 0 (0.0%) | 2 (33.3%) | 1 (16.7%) | 3 (50.0%) | 6 |
| Timeliness | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 2 (33.3%) | 4 (66.7%) | 6 |
| Comparability | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 4 (66.7%) | 1 (16.7%) | 6 |
| Usability | 0 (0.0%) | 0 (0.0%) | 1 (16.7%) | 3 (50.0%) | 2 (33.3%) | 6 |
| Relevance | 0 (0.0%) | 1 (16.7%) | 0 (0.0%) | 0 (0.0%) | 5 (83.3%) | 6 |
| Total | 0 (0.0%) | 2 (6.7%) | 3 (10.0%) | 10 (33.3%) | 15 (50.0%) | 30 (100.0%) |

**Table 3. Data quality dimension scores**

The characteristic level, dimension level, and overall ordinal ratings displayed in Table 1 were also summarized as continuous data that range from 1 (unknown) to 4 (appropriate). Again, note that where a category was scored as "not applicable (0)" the score was reset to null or missing (see Methods). In total, of the 24 data quality characteristics, about half were summarized as having a mean score of 3.5 ("halfway" between "marginal" and "appropriate") or higher and half below.

More specifically, for the evaluations available to date, under-coverage, item non-response, edit and imputation, processing, estimation, timeliness, comprehensiveness, integration, standardization, equivalency, adaptability, and value and had a mean score of greater than or equal to 3.5. Eight characteristics (i.e., over-coverage, collection and capture, linkage-ability, product comparability, historical comparability, accessibility, documentation, and interpretability) had mean scores that ranged from 3.0 (marginal) to less than 3.5 and the characteristics of unit non-response, reliability, correlated response variance, and simple response variance had means of 2.7, 2.3, 2.3, and 1.3, respectively. The standard deviations (STDs) ranged from 0.0 to 1.5.

While the dimensions of timeliness and relevance had mean scores of 3.5 or higher (i.e., mean timeliness=3.7 (STD=0.5) and mean relevance=3.5 (1.2)), both the dimensions of accuracy and usability received a mean score of 3.2 (STD = 1.0 and STD = 0.8, respectively) and comparability had a mean score of 2.8 (1.0). Lastly, of the six database evaluations available at the time of the study, three of the six databases were scored as "appropriate" overall, one was found to be "marginal", and two were rated "not acceptable" overall (Table 1). Summarized as a continuous variable, the mean overall score for the six databases was 3.2 (1.0).

# Discussion

The task of organizing, maintaining, and improving over 20 health administrative databases, many of which are person-oriented and population based is challenging. Implementation of a new strategy designed to facilitate database maintenance and data quality improvement for CIHI was initiated in April 2000. A sub-component of the new strategy includes a standard method for database data quality evaluation that appears to be unique in the field. The objectives of evaluation are to identify and rank aspects of data quality needing improvement and to assist in the documentation process. Prior to use, the evaluations must be validated. To date, the available results have not been completely validated. Although the results are not yet fully suitable for database improvement, they were presented solely for the purpose of demonstrating how the evaluation process might be of value to Senior Management or other decision makers in charge of numerous data holdings. Hence, the purpose of this study was to outline the CIHI database evaluation process and to present some early results based on the process.

One way of tracking the impact of the Quality Cycle or the performance for CIHI as a whole might be to track the number of databases assessed as "appropriate". To date, if taken at face value, the results suggest that overall data quality at CIHI tends to be good. The mean overall score for the six databases for the Institute as a whole was 3.2 (n=6, STD=1.0, Range=2 (not acceptable) to 4 (appropriate)), and the same result expressed as a categorical variable indicates that three of the six databases were scored as "appropriate" (4) overall, while one was found to be "marginal" (3), and two were rated "not acceptable" (2) overall (Table 1).

While a summary score is convenient, it is crude. For that reason, the data quality dimensions provide a more specific sense of what data quality concepts may be the main drivers of the relatively positive Institute score and what main concepts need attention. Again, when broken down by dimension it appears that the most common response category score was "appropriate" (15/30 or 50% (Table 3)). As well, database relevance and timeliness tended to be rated as "appropriate", however, database comparability and usability were flagged as areas in need of attention. If the findings are assumed valid, Senior Management can feel confident about the relevance and timeliness of the CIHI databases and, at the same time, they might direct resources to improve the comparability and usability of the data holdings.

Even more focused information of what to target and why, is available from the twenty-four data quality characteristics scores. Yet again, across the five response categories for the twenty-four characteristics of data quality, the most frequent ordinal scale score was "appropriate" (85/144 or 59%). However, "unknown" was also frequent (21/144 (15%)), which suggests that a fair amount of data quality characteristic information is undetermined. In particular, simple response variance, reliability, and correlated response variance are characteristics that tended to be identified as "unknown" for CIHI as a whole. Other flagged areas include coverage and response, as well as documentation and interpretability (Table 2).

Since the findings can be summarized for an institution as a whole, and not just by individual database, they can be used to direct data quality improvement initiatives for an institution as a whole. For instance, the early findings indicate that a corporate-wide initiative to improve health care institution tracking ("frame maintenance") could result in higher coverage and response scores and ultimately such an initiative might improve the accuracy of the Institute's databases. Statistics Canada has recently started a long-term frame project that might serve as a model or the responsibility of maintaining a single nation-wide frame database might be shared across the two organizations so that efforts are not duplicated. Similarly, corporate-wide initiatives to regularly obtain reliability/validity coefficients and to improve database documentation (e.g., easily accessible centrally located electronic documentation center) would also improve the Institute's accuracy and usability ratings, respectively, over time.

The results presented by each database suggest that databases 1, 4, and 6 have demonstrated "appropriate" levels of data quality overall and databases 2 and 5 are priority areas in need of resources for database improvement (Table 1). For database 2, while timeliness and relevance were found to be appropriate, database accuracy and usability were flagged for improvement. To understand what is causing the accuracy of database 2 to be low, the eleven characteristics can be considered. Based on the characteristics for database 2, it appears that more needs to be known about the reliability and validity of the data, and collection and capture, unit non-response, and edit and imputation should be improved in order to achieve greater accuracy.

Based on these promising initial results the evaluation process appears to have been successful in meeting the primary objective of identifying and ranking aspects of data quality requiring improvement. Not only does it highlight adequate areas or, conversely, target problematic areas within a database, it appears to facilitate the seemingly overwhelming task of understanding the state of data quality for numerous data holdings.

As demonstrated, the evaluation scores can also be used to describe data quality for an institution as a whole and it appears that strategic, corporate-wide planning might be facilitated when the evaluation results are summarized across the holdings. Moreover, the results might prove accessible to those with a wide range of expertise and are conducive to summary analyses. While some limited descriptive statistics were presented, the results produced by the evaluation process can be easily summarized with more graphically sophisticated techniques to aid in interpretation. It should also be noted that, even though the number of evaluations is still low, numerous database improvements have already been implemented because of the evaluation process and many of these improvements might not have been otherwise detected.

As far as meeting the second objective of facilitating the documentation of data quality for users, it is too early to tell. Given that the data quality document for users template is similar to the evaluation report template (except that evaluation scores and confidential data are removed) it is anticipated that the evaluations will help.

In terms of early insights, it goes without saying that before taking on an endeavor such as a corporate-wide database evaluation effort, the evaluation methodology should be based on the relevant literature and should be methodologically rigorous as well as practical [6]. Clear definitions should be provided and, if the evaluation process is to work, everything must be held as constant or as standard as possible. Analytical support, templates, and detailed step-by-step procedures are necessary. Moreover, the evaluation process itself should be subjected to regular quality improvement [8].

Though it is still early, it appears that it is crucial to have an objective outside party involved in the evaluation process. This observation has been made by both the database staff, who appreciate the objective perspective, and by the independent evaluators. One solution might be for the Data Quality Section to conduct the evaluations independently but in tandem with the database areas and then the results could be compared. Not only can outsiders offer a new perspective, they can also help assure that the data quality of a database is measured in as standard a way as possible and the database is fairly ranked in order to help ensure that resources for improvement are distributed according to need.

To facilitate the implementation of a new data quality strategy, it is essential that staff is assured that the scores are in no way a reflection of their personal abilities. It must be emphasized that the objective is to evaluate the databases in a clear and standard way for the purpose of equitable resource allocation for data quality improvement and to inform users of the limitations. To this end, and though there is clear “bottom-up” and “top down” support for the evaluation component of the new strategy, to augment implementation the process may undergo a name change from database “data quality evaluation” to “data quality improvement inventory”.

Lastly, to facilitate the implementation of a new data quality strategy it is important to convey the gravity of the possible impact of faulty data on peoples’ lives and the amount of “scrap and rework” that might be prevented if integral database maintenance and a thorough data quality analysis are conducted in the first place. It is crucial to illustrate the impact of faulty health data and instill a sense of duty towards the patient population. It must also be made very clear that people must inform users about the limitations of their data. If the limitations are unknown, then this must be clearly communicated. It is better to let users know that the quality of the data is unknown than to run the risk of letting them assume that the data are valid. For these reasons, the study and communication of the extent, impact, and resolution of data quality, and hence information quality, must be more forcefully pursued. Nowhere else may this be more pertinent than in the health care sector where critical decisions are being made and lives may be in the balance.

### *Costs*

Finally, the cost of data quality evaluation must be addressed. The importance of a cost-benefit analysis of data quality evaluation within CIHI has resulted in a recent effort to track relevant time and resource use for evaluation. However, examples of the costs or benefits expressed in monetary terms are not yet available. Although results are not available, some preliminary cost observations have been made. Outside of the cost of a new data quality unit responsible for the development and support of the database evaluation process, CIHI experience to date suggests that most of the resources, e.g., database managers, analysts, technical support, and documentation, required for database evaluation are already in place [8].

While it was estimated that execution of the Quality Cycle might take 30 person days across a database team, it appears that some areas have required 30 - 60 days and it has been noted that some products or services have been cut back so that implementation of the Quality Cycle, including database evaluation, could take place. Given that many areas are already conducting data quality activities, other than the initial study time required learning the revised standard approach, for the most part, it appears that no additional resources have been necessary. One preliminary observation, however, is that cost, as measured by the time required for existing staff to complete an evaluation, appears to be inversely related to the level of methodological expertise available from a database area [8].

In addition to existing data quality efforts, as mentioned, a new CIHI Data Quality Section (three methodologists, a classification expert, some administrative support, and a manager) has been put in place and is devoted to studying data quality issues, as well as to data quality evaluation methodological development and support. While replicating such a unit in many external health care settings for the purpose of database evaluation would be costly, there may be no need to do so.

Although the evaluation process has not yet been implemented externally, some additional observations based on our experience can be made. First and foremost, what is required for database data quality evaluation is methodological (i.e., epidemiological or statistical) expertise, as well as information systems, clinical, and coding professional expertise and many facilities already have such expertise in place. As well, Senior Management commitment and an assurance of resources in all operational plans for data quality improvement have proven crucial for successful database data quality evaluation at CIHI and should be considered for any external evaluation [8].

### *Limitations*

The main limitations are that the measurement properties of the evaluation process are not yet known and only preliminary data were available at the time of the study. To interpret the results with confidence, only complete and independently validated evaluations should be included for study and only like years of data should be combined. Algorithm overrides must also be kept to a minimum. As each area adopts the standard Quality Cycle, as outlined in the framework, and as the Data Quality Section becomes more familiar with each data holding, the independent verification process should become timelier.

## Conclusion

Although data quality has been a priority since the establishment of CIHI, a new method has been recently implemented to systematically evaluate and improve the data quality of the CIHI data holdings. Preliminary results suggest that the method is successful in its primary objective to identify and rank aspects of data quality requiring improvement. For those responsible for maintaining numerous data holdings, the results based on the method might make understanding the data quality status and relative needs of each of the databases much more accessible. This type of summary tool might prove useful for

not only improving data quality, but more importantly, for improving the information that is based on the data.

**Acknowledgement:** The authors thank Frank Ivis, Sandra Mitchell, André Lalonde, and Bruce Petrie for their thoughtful input and critical reading of the manuscript and for their support without which this paper would not be possible.

# References

[1] Brackstone, G. "Issues in the Use of Administrative Records for Statistical Purposes," *Survey Methodology*, 13, 1987.
[2] Brackstone, G. "Managing Data Quality in a Statistical Agency," *Survey Methodology*, 25(2), 1999
[3] *Data Quality Manual (Version 1)*, Canadian Institute for Health Information, 2001.
[4] Deming, W.E. *Out of the Crisis*, W. Edwards Deming Institute, 1986.
[5] Deming, W.E. *The New Economics for Industry, Government, and Education*, W. E. Deming Institute, 1994.
[6] Eppler, M. J. and Wittig, D. "Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years," *Proceedings of the 2000 International Conference on Information Quality*, MIT, 2000.
[7] Huang, K.T., Lee, W. L., and Wang, R. Y. *Quality Information and Knowledge*, Prentice-Hall, 1998.
[8] Long, J., Richards, J., and Seko, C. "The Canadian Institute for Health Information (CIHI) Data Quality Framework, Version 1: A Meta-Evaluation and Future Directions," *Proceedings of the Sixth International Conference on Information Quality*, MIT, 2001.
[9] *National Consensus Conference on Population Health Indicators*, Canadian Institute for Health Information, 1999.
[10] *Products and Services Catalogue*, Canadian Institute for Health Information, 2002.
[11] Richards, J., Brown A., and Homan C. "The Data Quality Study of the Canadian Discharge Abstract Database," *Proceedings of Statistics Canada Symposium, Achieving Data Quality in a Statistical Agency: a Methodological Perspective*, 2001.
[12] *Statistics Canada Quality Guidelines, 3rd Edition*, Statistics Canada, 1998.
[13] Strong, D. M., Lee, Y.W., and Wang, R.Y. "10 Potholes in the Road to Information Quality," *Computer IEEE*, 1997.
[14] *Statistical Policy Working Paper 4 – Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics*, U.S Federal Committee on Statistical Methodology, 1987.
[15] www.cihi.ca