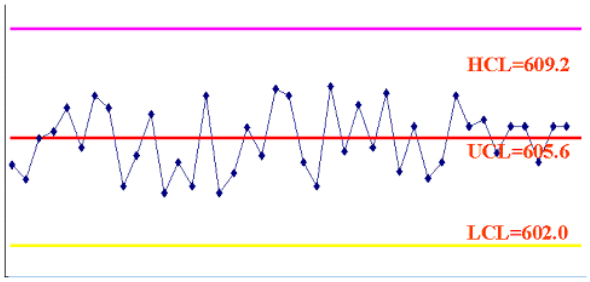
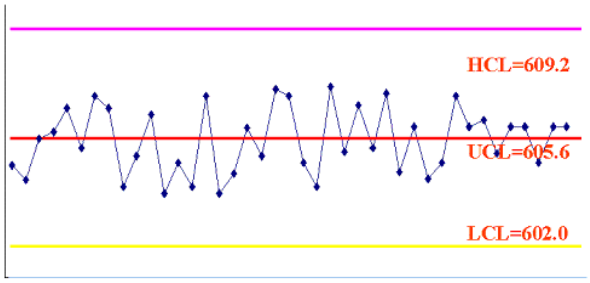
Georgia Division of Public Health

Gold Standard
Data Quality Protocol



Georgia Division of Public Health

Gold Standard
Data Quality Protocol



May 2002
2nd Edition

Georgia Department of Human Resources
Division of Public Health
Office of Health Information and Policy

Prepared by:
Jacqueline R. Bennett-Burkholder, Ph.D., M.S. and Gordon R. Freymann, M.P.H.

Architect: Frank H. Millard, M.A.

With contributions from
Teresa C. Lofton, Ph.D., M.P.H.

Suggested citation: Bennett-Burkholder, J.R., Freymann, G.R. (2002). Georgia Division of Public Health: Gold Standard Data Quality Protocol. Atlanta: Georgia Division of Public Health (publication # DPH02-111)

Georgia Department of Human Resources
Division of Public Health
Office of Health Information and Policy

GEORGIA DIVISION OF PUBLIC HEALTH
OFFICE OF HEALTH INFORMATION AND POLICY
GOLD STANDARD DATA QUALITY PROTOCOL, VERSION 2.0

Table of Contents

Executive Summary

Performing each step of the Georgia Division of Public Health (GDPH) Gold Standard Data Quality Protocol assures the quality of all Divisional data is in a known state. The methods provided in this document provide the means to ensure Division data are of an acceptable level of quality. Use of the procedures outlined in this document enable data users to compare variables on data quality dimensions such as accuracy and completeness.

This document will enable the reader to:

1. Understand how to read properties of a data item (aka "variable," or "field"),
2. Define and document the properties of a data item,
3. Assess whether data items are aligned with standard data properties,
4. Use appropriate methods to assess the degree of each data item's quality, as measured by the number and percent of invalid values,
5. Determine which data items are suitable for further error correction,
6. Develop rules that identify data errors, and derive-and-store valid values where errors are found, and
7. Prepare a data quality audit report.

Such efforts support the implementation of policies for standard methods of data collection, management, reporting and distribution established by the GDPH Data Policy Committee. Quality information of all Division data assets is a prerequisite for valid and reliable data analysis and interpretation that create evidence about the health status of Georgians. This assessment function is core to the Division's business, and is required to know the degree to which its policies and actions maximize health status.

II. Introduction

Purpose

Data quality is defined as the measurement of the ability of data to meet the needs and requirements of the user. Data quality measurement is multi-dimensional: quality is measured with respect to dimensions such as accessibility, accuracy, completeness, relevancy, and timeliness. The work of the Massachusetts Institute of Technology Total Data Quality Management Group on Data Quality, Knowledge and Information Management, is the Georgia Division of Public Health's main reference for data and information quality development. The group advocates this multi-dimensional approach to data quality auditing, and the concept that information is managed as a product of core business functions rather than a byproduct. The methods described in this protocol address the following data quality dimensions:

1. First, ***Representational Consistency*** – Methods are defined to identify the properties of data items and to correct data properties that are not represented by or aligned with the standard format.

2. Second, ***Accuracy*** – Methods are defined to identify data errors and to document invalid values; and to correct errors in order to assure the reliability of the data and the identification and correction procedures.

3. Lastly***, Completeness*** – Methods are defined to set criteria for data completeness, and to identify data that are not in a known state such that a state of data completeness can be achieved.

III. Data Quality Requirements

III.A. Data Management Requirements

As a prerequisite to conducting a data quality audit, (an automated assessment of the degree to which data adheres to established rules and specifications), all data must be compliant with GDPH *Data Quality Requirements*, detailed below. These requirements serve as a reference for all data quality management. Definitions of terms used in these requirements can be found in Appendix A.

1. ***A variable shall have one and only one name***. Example: the data item "sex" is named "sex" as opposed to "gender." Or, if a database collects information about "permits," then that construct shall not also be referred to as a "certification."
2. ***A variable shall have one and only one definition***. Example: The definition of white race is:

   > A person having origins in any of the original peoples of Europe, the Middle East or North Africa.

3. ***A variable shall be stored as one and only one data type***. For example, string data such as ICD codes (cause of death codes with values such as A018, 001.1) should not be stored as a numeric field.
4. ***A variable shall have one and only one field length***. Example: collection of street address should allow for 40 characters, not less.
5. ***A variable shall be stored in one and only one unit of measurement***. Example: Birthweight is stored in grams, but not pounds and ounces.
6. ***A variable shall be stored in one and only one level of measurement***. Example: a *nominal* variable such as "race" shall not be stored as *interval* data.

7. ***A variable shall be represented by or stored as only those values specified in its definition***. Example: if 1=yes and 0=no, there should be no other values (e.g., 9, 8, z, abc) found in that field.
8. ***Data objects shall have one and only one source***. For example: official Georgia Birth statistics will come from the Office of Health Information and Policy, Georgia Division of Public Health.
9. ***No duplicate sources of data objects (storage or collection) shall exist***. Example: A central data repository for analytic health information shall be established to contain each data domain (such as vital records, notifiable diseases, immunizations).
10. ***No duplicate records shall exist in data objects***. Example: One and only one record for each birth in the birth data domain.
11. ***All data domains, data objects and variables shall be free of data anomalies.*** All data assets will be examined for invalid values and such values will be processed such that their values are in a known state.
12. ***Unknown, missing and inapplicable values shall be represented by one known value***. Example: 99 (unknown), nulls, blanks, or out of range values all set to = -1.
13. ***Unknown, missing and invalid values in all data domains shall have consistent representations as defined in 12.*** .
14. ***All data dictionaries shall define the following data properties for each variable.*** The properties are in the following table:

Variable Name(s)	Name of the data item used for storage and if applicable, presentation. Storage names begin with a domain identifier (e.g. Birth), followed by an owner (e.g. Mother), followed by the variable name (e.g. birth.mother.education_level). Presentation name (or label) is used to present data, such as "Mother's education level."
Definition and Variable Associated Standards	A statement containing the reason to collect or use the variable, and external standards that apply to the variable.
Valid Values	Acceptable values for the variable being defined (e.g. mother's age range = 10-55 years inclusive.).

Data Type	The characteristic of a variable that determines what kind of data it can hold. For example, data types include Byte, Boolean, Integer, Long Integer, Currency, Decimal, String, Double, and Date.
Field Length	The number of numerical places or characters for a specific field.
Unit of Measurement	(See glossary for definition)
Level of Measurement	(See glossary for definition)
Unit of Analysis	The unit of measurement assigned to a variable for analysis.
Level of Analysis	The level of measurement assigned to a variable for analysis.
Derivation	For calculated fields, the variables used and method to derive the calculated variable.
Time Stamp of Standard	The date on which the variable definition was defined or revised.

An example for the variable "infant's sex":

Property	**Value**
Presentation Name	INFANT'S SEX
Storage Name	BIRTH.INFANT.SEX
Definition	Biological sex of the infant at delivery.
Valid Values	1=Male, 2=Female, -1=unknown.
Data Type	Integer
Field Length	2
Unit of Measurement	Unitless
Level of Measurement	Nominal
Unit of Analysis	Unitless
Level of Analysis	Nominal
Derivation	N/A
Time Stamp of Standard	5/4/2000

If a data property is not applicable for a variable, "N/A" shall be noted.

15. ***The following notation shall be used when describing a variable in terms of class and object: Class.Object.Variable*** – such as, *VitalRecords.Birth.Birthweight*. Further, a variable's notation will show the standard variable whose specifications are followed: "VitalRecords. Birth.Mother_Race.Race" says "of the vital records class of data, birth records, there is a variable stored as 'mother_race' that inherits its specifications from the standard variable 'race.' "

III.B. Data Quality Audits

Data quality audits refer to the process of assessing the quality of values contained in each data item (field).

1. ***All domains shall have data quality audits***.
2. Data quality audits shall include:
    - 2.1. ***Range checking of integer and real numbers*** (e.g. if 'age' can range from 0-120 years, then a value of 122 is invalid).
    - 2.2. ***Value checking of variable contents***.
    - 2.3. ***Pattern checking of strings and dates*** (e.g. values for a date field that are always the same ('date heaping') may indicate error).
    - 2.4. ***Functional and logical dependency checking***.
    - 2.5. ***Logical constraint checking within variables, records and objects*** (e.g. men should not be recorded as having cervical cancer).
    - 2.6. ***Inexact (range) constraint checking***.
    - 2.7. ***Statistical (outlier) constraints checking***.
    - 2.8. ***Check for the correct treatment of unknown, missing, and inapplicable values*** (e.g. "999" translated into –1).
    - 2.9. ***Check for unlikely probabilities of combinations of variables*** (e.g. birthweight of 3200 grams and gestational age of 24 weeks).
3. ***Data quality audits shall be ongoing as data are added or modified***.
4. ***Procedures and policies shall be in use to direct and assure data cleaning meets these data quality requirements***.
5. ***The process of data quality auditing, anomaly detection, data cleaning shall be automated***.

IV. Steps to Assess Alignment with Standard Data Properties

Step 1: Documentation of Data Properties

The primary step in assessing the degree to which the data collection adheres to the data management requirements is to document the status of each requirement. Document the following data properties from the data dictionary, file structure, or coding specifications for each data variable (as shown in section III.A, #14):

> *Presentation Name, Storage Name, Definition, Valid Values, Data Type,*
>
> *Field Length, Unit of Measurement, Level of Measurement, Unit of Analysis, Level of Analysis, Unknown, missing, invalid Value Notation, Derivation, Time Stamp of Standard, and Date of Assessment.*

Step 2: Corrective Actions

If any of the data item's properties do not align with the standard, then corrective action is in order. For example, if the data item's name, definition, data type, or field length do not meet the standard, then modify the property to align with the standard. In the case that standard data properties are not yet defined, submit a proposed set of data item properties for review by the Data Policy Committee.

Currently, data property standards are available at
http://health.state.ga.us/healthdata/datause.shtml

V. Steps to Develop Data Error Detection Procedures from Descriptive Statistics

Descriptive statistics including frequency distributions, measures of central tendency (mean, median, mode) and measures of dispersion (range, standard deviation, percentiles) are used in this protocol to summarize patterns and trends in data.

Step 1: Data Extraction

Extract a dataset from the database. If possible, extract a minimum of 5 years for historical comparisons.

Step 2: Frequency Distribution

Run a frequency distribution for each data item. Because of the quantity of output, characteristically this is only done for variables with less than 100 unique values for each year and each domain. Table 1 is an example of a frequency distribution of the data item 'race.'

	1998		1999		2000		Table Total .00	
	Count	Col %	Count	Col %	Count	Col %	Count	Col %
-1 Unknown	*22890*	*18.3%*	*23123*	*17.9%*	*25077*	*18.6%*	*155503*	*18.2%*
1 White	*71659*	*57.3%*	*73872*	*57.2%*	*75553*	*55.9%*	*489895*	*57.3%*
2 Black or African-American	*28023*	*22.4%*	*29347*	*22.7%*	*30967*	*22.9%*	*193303*	*22.6%*
3 Asian	*2159*	*1.7%*	*2537*	*2.0%*	*2949*	*2.2%*	*14705*	*1.7%*
4 American Indian or Alaska Native	*255*	*.2%*	*234*	*.2%*	*299*	*.2%*	*1524*	*.2%*
5 Native Hawaiian or Other Pacific Islander	*30*	*.0%*	*28*	*.0%*	*27*	*.0%*	*192*	*.0%*
6 Multiracial	*13*	*.0%*	*107*	*.1%*	*186*	*.1%*	*321*	*.0%*
Table Total	*125029*	*100.0%*	*129248*	*100.0%*	*135058*	*100.0%*	*855443*	*100.0%*

Table 1 (fictitious data: years will not equal total)

Note that in Table 1 all records have a valid race value (1-6) or an acceptable unknown value (-1), and valid names. Also note that the number of values

remains within expected limits – that is, the number of Asian births each year was 2,159, 2,537, 2,949 – a steady increase without for instance drastic drops in numbers.

Step 3: Histograms and Statistical Outliers

Continuous data should also be viewed using histograms and statistical analysis. Histograms provide a visual representation of the data. Statistical analysis of the mean, median, mode and upper and lower percentiles also assist in interpreting the information.

A histogram of the values contained within a data item is very useful. For example, the following histogram of gestational age quickly reveals that '99' is contained in the data item (often '99' or '9' or '999' is intended to mean 'unknown' in source data) (Figure 1).

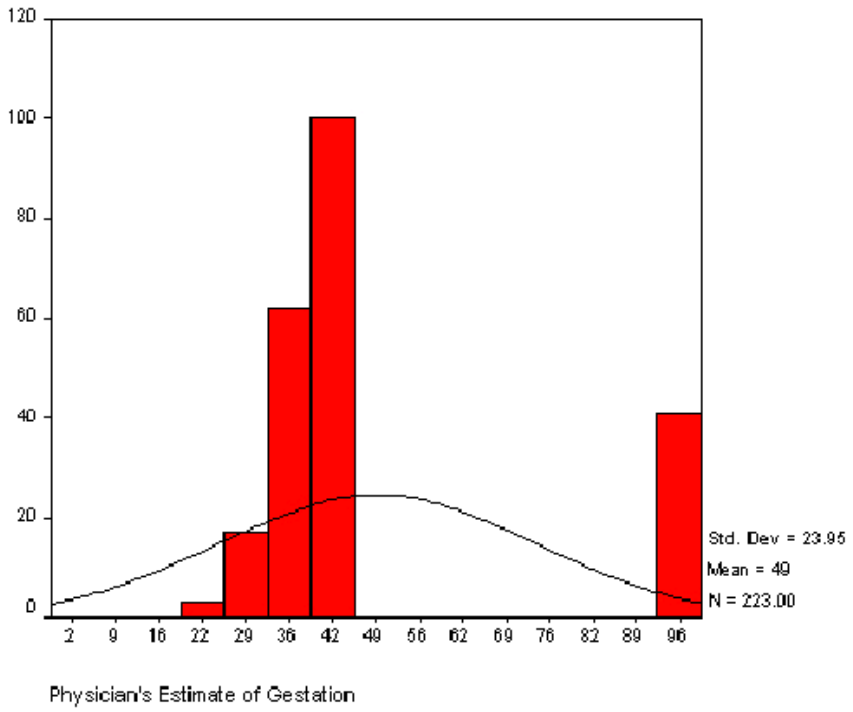


Figure 1

It is important to obtain this view before performing statistical analyses for continuous data. Statistical analysis for continuous (non-discrete) variables should include:

1. Mean
2. Median
3. Mode
4. Standard Deviation
5. Range (minimum, maximum)
6. 0.05 and 99.95 Percentiles (potential outliers)

Table 2. Gestational Age in Weeks

N	Valid	*135037*
	Missing	*21*
Mean		*38.73*
Median		*39.00*
Mode		*40*
Std. Deviation		*3.771*
Minimum		*12*
Maximum		*99*
Percentiles	.05	*20.00*
	99.95	*99.00*

Table 3. Gestational Age in Weeks

AFTER removing '99's'

N	Valid	*134712*
	Missing	*0*
Mean		*38.59*
Median		*39.00*
Mode		*40*
Std. Deviation		*2.339*
Minimum		*12*
Maximum		*44*
Percentile	.05	*20.00*
	99.95	*43.00*

Table 2 provides an example of statistical analysis of the data item 'gestational age' without accounting for '99's' discovered via the histogram.

As presented in Table 2, the minimum value for physician estimation of gestation is 12 and the upper value is 99; However as previously stated, in this case unknowns were coded as "99" at the source. Therefore this data item is not in alignment with the requirement that unknowns be coded with a –1. The standard deviation and maximum in Table 3 reflect the removal of "99's."

Table 3 allows us to verify if other invalid values exist. If the

standard range for this data item is 20-44, then we see by the minimum of '12' that this data item does not align with standard properties.

Revisiting the histogram prepared in Figure 1, 99's are excluded and an alternate view of this data item's values is possible (Figure 2).

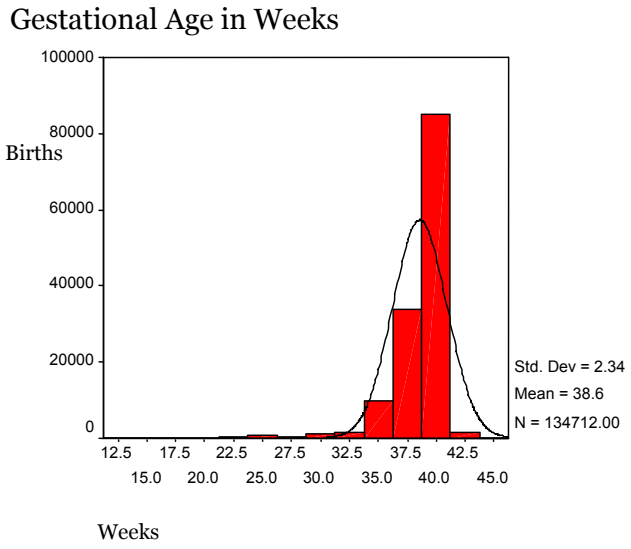


Figure 2

Therefore, by using these techniques (histograms, frequency distributions, cross-tabulations (a frequency distribution of two data items together), obtaining the range, obtaining the outliers, etc.), one can develop appropriate rules (protocol; set of procedures) for each data item to detect the following types of data errors: ***Unknown*** (Missing, null) values, ***other invalid values*** (e.g. out of range values), ***invalid strings*** (that which is not included in the variable's domain specifications. For example, the acceptable specifications may be "yes" or "no." The value "false" would be considered an invalid string), ***invalid dates*** (that which is not included in the variable's domain specifications. For example, a date of birth occurring in the future), ***logical constraint relationships*** (exists if a value of a dependent variable exists only if a particular value of an independent variable exists within a record. For example, the state of being pregnant exists

only in females (not males)), ***functional and logical dependency relationships*** (exists when a value of a variable is derived from the value of another variable. For example, when unknown, gestational age can be estimated from birthweight. Therefore a functional and logical dependency check can be initiated for gestational age based on its relationship with birthweight: infants that weigh more than 3,000 grams should not also have a gestational age less than 20 weeks), and ***unlikely combinations of values between variables*** (exists when specific values of two variables are unlikely for a single record. For example, the combination of the variable values Mother's Age = 14 and number of previous live births = 2 is unlikely).

VI. Steps to take when Invalid Values are Found

1. Generally, if data are current, send back to the data source for correction, otherwise:

2. Proceed with the next section.

VI.A. Methods to Derive and Store Valid Values

When invalid values are discovered (including unknowns), a standardized process may be established in order to derive and store valid values where invalid ones exist. The procedures one follows to derive and store such values are represented by Activity Diagrams. A visual representation of the information flow that documents the decision process for invalid values is shown in Appendix B.

Before defining such rules, a limit must be set on the percentage of records that could be subject to additional derive-and-store rules. The ***5% Rule*** states that the upper limit is five percent; therefore, data item errors occurring in more than five percent of records are set to –1 (unknown) and left alone. For example, if for a given data item in a database, the percent of records that are unknown and missing is 4%, and the percent of records that contain invalid values is 3%, all such values for those records are set to –1 and no further action is taken. If however, the total of unknown and invalid (which in essence are the same concept: invalid values are for all practical purposes unknown) values is less than 5%, the data item containing these unknown values is a candidate for further derive-and-store procedures.

VI.B. Further Criteria for Error-handling Rule Development

In addition to the 5% rule mentioned in the previous section, one should ask whether the data item in question has business value. Basically, business value can be determined by asking yourself these questions:

- Once I've collected this data item, how will I use it?
- Is the data item critical to assessing business performance?
- Is the data item critical to monitoring business processes?
- To what extent will the data item in question be used as an input for a subsequent calculation or derived variable that is of business value? Subsequent derived variables will be affected by unknown values in the source data item.
- (and maybe) How frequently is the variable used?

The preceding questions assume one has a precise understanding of the mission and business of the organization in which one works.

VI.C. Derive and Store Rule Template

When developing rules, the following template is used (Table 4). The template is populated with information about the data item 'infant's sex.' As shown, 30 records had invalid values (in this case, all invalid values were "9's"). Of over 800,000 records, the percentage that are invalid is less than 1%, so the 5% Rule is not broken. A derive-and-store rule was developed, and documented using the template.

Table 4: Error-handling in the data item Infant's Sex

Invalid value rule: Any value other than 1 or 2.				
# Invalid	# Total Records	% Invalid	Derive-and-Store Rule	Reference
30	853,753	0.00%	If last digit of certificate number is odd, INFANT'S SEX is assigned female (2); if last digit of certificate number is even, Infant's Sex is assigned male (1).	NCHS[1]

Testing Rules: Rules should be designed to be complete. That is, they should handle all instances of invalid values. For example, the following rule template is not sufficient:

Table 5: Error-handing in the data item Mother's Race

Invalid value rule: Records with Mother's Race not equal to 1 through 6.				
# Invalid	# Total Records	% Invalid	Derive-and-Store Rule	Reference
12	853,753	0.00%	Store in Mother's Race the value found in Father's Race.	- -

What Table 5 assumes is that for each record where mother's race is not valid, a valid value will be present for father's race. This is not the case and shows the need for **testing** rules for completeness.

VI.D. Other Types of Derive and Store Rules

More complex rules, that for example use logical constraint relationships between variables, can be developed by searching relevant literature and meeting with subject matter experts. Such rules use a cross-tab (2x2 table) to document the relationship.

For example, Table 6 documents error based on a functional and logical dependency relationship between birthweight and gestational age:

[1] Instruction manual part 11: Computer Edits for Mortality Data, Including Separate Section for Fetal Deaths Effective 2000

Table 6: Infants Birthweight In Grams > 3000 & Gestation Weeks < 20

		Gestation Weeks			Total
		14	17	18	
Infants Birthweight In Grams	3005			*1*	*1*
	3010	*1*			*1*
	3033	*1*			*1*
	3090			*1*	*1*
	3260		*1*		*1*
	3515			*1*	*1*
Total		*2*	*1*	*3*	*6*

Per subject matter experts, infant's with a birthweight of 3000 or more grams do not have gestational ages of 14, 17, or 18 weeks; and vice versa. Therefore the following rule in Table 7 might be appropriate.

Table 7: Error-handling for the data items Birthweight and Gestational Age

Logical constraint rule: Records where BIRTHWEIGHT ≥ 3000 and GESTATION < 20 weeks.				
# Invalid	# Total Records	% Invalid	Derive-and-Store Rule	Reference
6	853,753	0.00%	Impute infant's gestation to the median value found in other records of the same birthweight.	- -

Note that the rule in Table 7 assumes birthweight is more likely to be correct than gestational age. This may be a correct assumption, based on prior evidence gathered during the data error detection phase described in section V. Note then too, that the order in which derive-and-store procedures are executed becomes critical where dependency relationships exist.

VII. Data Quality Audit Summary Report

In order to track rules that have been applied to data, it is imperative to document this information. The following templates display the requirements for documentation (tables 8 and 9):

Table 8: *Data Quality Audit Summary Report (A)*

Information Documentation	Example
Data Source	Vital Records 2000 Birth File
Number of Records	135,989
Number of Variables	147
Reviewed by:	Yours truly
Date Revised:	5/6/02

Table 9: *Data Quality Audit Summary Report (B)*

Data Item Reviewed	Number of Invalid Values	Percent of Records	Derive-and-Store Rule(s)	Number of Invalid Values now Valid
Infant's Sex	13	0.01%	If last digit of certificate number is odd, INFANT'S SEX is assigned female (2); if last digit of certificate number is even, Infant's Sex is assigned male (1).	13
Data Item 2	n	n.nn%	Rule….	n
Data Item 3 etc…	n	n.nn%	Rule…	n

LAST BUT NOT LEAST

Note the last column in Data Quality Audit Summary Report (B): Number of (formerly) invalid values that are now valid. To know this, the data error detection methods described in section V **must be run again on the data** to assure that the imputation rules were applied correctly. This action reflects the basic principle of a) defining your output, b) implementing a method to arrive at your desired output, and then c) testing your output to determine if your output is valid.

Appendix A: Data Quality Requirements Definitions

These definitions shall be used in creating data quality requirements, any and all documents and policies and procedures relating to data quality management, and all formal discussions relating to data quality management. These definitions shall be the working language of data quality, data integration, data management, and data analysis within GDPH.

Accessibility – a dimension of data quality. Refers to data being available and the speed of access.

Accuracy – a dimension of data quality. Refers to data that are error-free, reliable; that errors can be easily identified; that data have integrity; and data are representative of their design (or standard).

Class – an abstraction of concepts that have the same properties and behavior; such as *vital records*.

Completeness – presented data have wide enough scope and depth to support required decision-making.

Concept – the ideas or notions that lead to the collection of observations and measurements, or that represent descriptions of phenomena; such as the notion that births or deaths need to be recorded.

Continuous Variable – a measurement variable that can assume an infinite number of values between any two fixed points, such as *birthweight*.

Data Anomaly – an exception to a domain specification or variable definitions that indicate errors in the data or failures in the rules contained in domain specifications. If an anomaly is due to a data error, then the data object is corrected. If an anomaly is due to a rule or standard failure, where the domain specification does not accommodate observed data, then the rules or standards should be corrected. For example, a mother's age value of 4 years is a data anomaly because "4" is not in the valid range of values.

Data Collection – a set or class of data items that are contained within a single file or table; such as *births* or *deaths*.

Data Dictionary – a single authoritative documentation of variable data properties.

Data Item – a single collected observation such as *birthweight* or *name*.

Data Property – a required element stored in a data dictionary that defines a variable's storage and analysis specifications such as variable *storage name* and variable *definition*.

Data Type –the characteristic of a variable that determines what kind of data it can hold. Data types include Byte, Boolean, Integer, Long, Currency, Decimal, Single, Double, Date, String. For example, *string* is the GDPH standard data type for the variable *street address*.

Derived Variables – a variable that is a function of two or more independently measured variables whose relations are expressed in a certain way, such as ratios, percentages, indices, and rates. *Infant mortality rate* represented by the total number of infant deaths occurring during a specified time period per 1,000 total live births during the same time period is a derived variable; as is *age* (vs. stated age), which is derived (calculated) from date of birth and an event date.

Discontinuous Variable – a measurement variable that can assume only fixed numeric values, with no intermediate values, such as counts like the number of hours a client waited for service measured in hours represented by 1, 2, 3, 4 etc.

Domain – a selected area of interest that contains a collection of objects that are instances of the domain specification. Domains can be observations, populations, measurements, or variables; *vital records* is a domain.

Domain Specification – the collection of concepts that apply to a domain; NCHS coding standards and rules are domain specifications.

Independent and Dependent Variables – if F is function or decision space, such that d = F (I), then I is the independent variable and d is the dependent variable, since a value of d is dependent on a value of I. For example, fetal weight (d) is a function of gestational age (I).

Imputation Rule – a type of method to derive and store a valid value. Imputation implies that a variable's value is derived from another record, or a value from a sub-population of other records, rather than from the record itself.

Individual Observations – observations or measurements taken on the smallest sampling unit such as the observation of an infant's birthweight, with the sampling unit being the infant.

Inexact range constraint – an inexact range constraint relationship exists when an inexact range (e.g. "less than or equal to", "not equal to") of values of a variable is dependent on an inexact range of values of another variable. For example, the number of live births a woman has had in her lifetime is dependent

upon the mother's age, but the relationship is not exact. For example, "if mother's age <=15 then the number of live births cannot be > 5."

Invalid Values – a value that is not included in a variable's domain specifications. For example, if valid values for a data item are 1-10, then an "a" or a "22" are invalid.

Level of Analysis – the *level of measurement* assigned to a variable for analysis.

Level of Measurement –

(1) Nominal: One data element is identified as disparate from another but no direction is implied. Categorical properties or labels; such as the variable race represented by White, Black or African-American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander; or Male/Female. Appropriate descriptive statistics are the mode (most commonly occurring value) and frequency counts;

(2) Ordinal: Variables can be ranked to show one value is more or less than another value; however the difference cannot be calculated, such as 'high/medium/low." Objects are ordered by some nominal category irrespective of magnitude, and irrespective to the distance between ordered levels; such as variable representing the level of agreement with a statement represented by disagree, somewhat agree, agree, disagree. Appropriate statistics are the mode, frequency, median (middle value), and percentiles.

(3) Interval: Variables have an established identical distance between them on a scale; however they do not have a true zero, such as grams, inches, days. Ordering of objects is respective to a nominal category, the distance between objects respective to the nominal category, and without respect to the magnitude of the nominal category such as the number of hours a client waited for service measured in hours represented by 1, 2, 3, 4 etc. Appropriate statistics are the mean (average), median, mode, standard deviation (square root of the variance), range (maximum value – minimum value) and percentiles.

(4) Ratio: Have a true zero. Objects are ordered respective to a nominal category, where the distance between objects is known, and each objects measurement is respective to a known zero value such as annual income measured in dollars represented by 20,051, 55,987, 42,042, etc.

Measurement Variables – variables whose differing states can be expressed in a numerically ordered fashion such sex represented by 1 = Male and 2 = Female.

Missing Value – occurs in a variable (data item) when there is no entry for that variable in a record. The standard representation of missing values (a type of unknown value) is -1. Missing values may or may not be valid.

Object – anything to which a concept applies; an instance of a concept; such as, *births* or *deaths*.

Population – the totality of individual observations about which inferences are to be made, existing anywhere in a given universe of interest, and definitely specified by space and time such as the total number of births in Georgia in the year 1999.

Quality Assurance - all the planned and systematic activities implemented within the quality system that can be demonstrated to provide confidence that a product or service will fulfill requirements for quality.

Quality Control – the operational techniques and activities used to fulfill requirements for data quality.

Range – a value that is within an acceptable range is a value that is below the maximum or above the minimum value specified in the variable's domain specifications. For example, if age ranges from 0-120, an age of 65 is acceptable. A value that is outside of an acceptable range is a value that is below the minimum or above the maximum value that is specified in the variable's domain specifications. In the same example, an age of 121 is unacceptable.

Ranked Variables – a variable that indicates order in observations without any assumption placed on the magnitude between ranks such as order in which a group of 5 children were immunized, represented by 1, 2, 3, 4, etc.

Record – a single instance of an object; one birth or one death.

Relevancy – data are applicable; relevant; usable; needed.

Reliability – the closeness of repeated measurements to the same value, or a process with inputs of equal value, will always produce outputs of equal value; also known as precision.

Representational Consistency – data are consistently represented in the same format and compatible with previous data.

Sample Observations – a collection of individual observations, from a population, selected by a specific procedure or program such as observations on a group of infants.

Statistical Outlier – operationally defined as the existence of data which falls above the 99.95th percentile or below .05th percentile.

Timeliness – refers to the age of data and whether data are up-to-date.

Transform data – refers to changing the data type, such as from string to integer.

Translate data – refers to changing the representation, but not meaning. Just like one would translate from Spanish to English, one might translate county codes from Georgia County Codes to FIPS codes: the representation is changed but the meaning remains the same.

Unit of Analysis – the unit of measurement assigned to a variable for analysis.

Unit of Measurement – (a) refers to the system of measurement: English or metric (CGS and MKS) or SI; (b) the specific unit, within a measurement system, at which measurements for a variable are made such as grams.

Unknown value – includes missing, null, and invalid values because in all such instances the value is unknown. The standard representation of an unknown value is -1. Unknown values may or may not be valid.

Validity – the closeness of a measured or computed value to its "true" value, or a process that measures exactly one and only one specific phenomena; also known as accuracy.

Variable – the actual property measured by the individual observations, and indicates the degree to which observations in a sample differ such as birth weight, race, or sex.

Variable Definition – a statement containing the reason to collect or use the variable, domains that contain the variable, any external standards that apply to the variable, valid values for the variable, and the method of derivation for the variable if it is derived.

Variable Field Length – the number of numerical places or characters for a specific field.

Variable Name – the name of the variable used for presentation.

Variable Precision – the level of granularity for variables with a ratio level of measurement (includes derived variables) represented by its number of decimal places.

Variable Storage Name – the variable's field or column name as it would appear in a database. Such names are typically not viewable by information consumers, but only database managers.

Variable Time Stamp – the date a variable was specified.

Variate – a single reading, score or observation or a given variable. For example "Asian" is a variate of the variable *race*.

Appendix B: Example of Imputation Decision Process: Mother's Race Data Item