



**Data Format and Variable Standard for GenomEUtwin's
Phenotype Database**

Version: 4.0

Ann Björklund
Jan-Eric Litton

Table of contents

Aim	3
Background	3
Introduction	3
Updates	3
EUid number	4
Meta data.....	5
Variable names	5
Data format.....	7
Appendix.....	13

Aim

This document describes the basic phenotype data and its variables that are stored in the GenomEUtwins databases.

The unique identification number for the individuals, EUID number, is explained.

How the data is stored in the GenomEUtwins centres databases is NOT described in this document.

Background

Database infrastructure has become a critical component for competing in life sciences research and discovery. The explosion of data requires that the data will be properly loaded, accessed, managed, queried, analyzed, and shared with others.

There is a lot of twin-data, and the amount is increasing every year. Much would be gained by standardising some aspects of the GenomEUtwin data handling. One way to simplify the time-consuming work of data management is to give the data a standardised, and thus well-known, format.

Introduction

In November 2002 decided the Database Core to build a prototype of a common database. To success with that goal was it absolutely necessary to have the same variable name and data format on the same variables. For that reason was the first version of this document released. The document also described how to deliver the data to Stockholm, where the prototype database was built.

In December 2004, about two years later, the GenomEUtwin centres have agreed to no longer send the data to Stockholm. Instead, each centre will store the data on their own database servers located on their own centre. The database servers will then be connected to each other by IBM's Personal Integrator.

For that reason has the information about the data transfer been omitted because there is no need for it any more.

Updates

Since version 3.2.4

The whole document has been rebuilt and conformed after the new condition to not send data to Stockholm.

Since version 3.2.3

The word "prototype" is no longer in the title.

Information about appendix is added.

The name of chapter 5. *File specification* is changed to 5. Data specification

Section 5.1 Other phenotypes is new and the following sections are renumbered.

Section 5.5 File format is renamed to 5.5 Data format.

Chapter 6 and 7 is merged and renamed to 6. Transferring Data.

The order of 3. Updates is changed. The newest changes come first.

EUid number

The EUid number is an identification number for the individuals whose data is stored in the database. The number is unique and each centre is responsible to create a EUid number for their individuals.

The EUid number consists of four parts: country, randomized number, identification number and check sum.

Country code

The country codes, giving data origin will be according to the ISO 3166 standard.

036 – Australia
208 – Denmark
246 – Finland
380 – Italy
528 – Netherlands
578 – Norway
752 – Sweden
826 – UK
999 – Unknown

Randomized number

In the database there will be twins and non twins. Each twin pair will share the same randomized number, a none twin will receive an own randomized number. The none twin randomized number will only occurrence once, but the randomized number for twins, will occur two times if twins, three times if triplets and four times if quadruplets.

The country code is part of the EUid number, and this allows each country to administrate their own randomized numbers. How this will be done is up to each country as long as it generates a unique number for each individual and contains of 7 numbers. The randomized number does not need to be generated in random, it can for example be derived from a locally used twin pair number.

Identification number

The EUid number needs indicator weather or not this is a twin. To obtain this information and still have unique numbers the EUid number should end (except the checksum) with the following numbers:

1 – Twin 1
2 - Twin 2
3 - Triplet
4 - Quadruplet

The indicator for individuals being none twins will be:

0 - Non twins

Checksum

The checksum is calculated by GUMM algorithm (H. Peter Gumm: A new class of check digit methods for arbitrary number systems, IEEE Transactions on information theory, 31 (1985), 102—105).

The GenomEUtwin Data Transfer Java application, <http://www.ktl.fi/morgam/apps/genomeutwin/>, can be used to calculate GUMM check digits.

Example

If there is a twin from Sweden with randomized number 0000212, idnum 1 and checksum 0 will the EUid number be:

752000021210

Meta data

Missing data

Several codes are used to code missing data. The different types of missing data is described below:

Irrelevant, non-participant, non-response Used when data is irrelevant for a person either due to not included in study base or due to non response to questionnaire/interview as a whole.

Irrelevant, structural Used when data is irrelevant in the context.

Do not know Used when the respondent explicitly states it.

Unknown Used when no answer is given to specific item, or when no data is available.

Accuracy codes for dates

At least some of the dates variables will probably have missing data according to some of the four reasons above. The easiest way to code the missing data is to use sequences of 9's, 8's etc. But the problem is that most database systems don't take these codes as valid. It's impossible to insert such a code in the database. To avoid this problem is an accuracy code variable added after each date variable.

For example:

The variable WEIGHT_SELF_DATE (Self reported weight – date) has an accuracy variable named WEIGHT_SELF_DATE_A.

If you have a complete date for this variable, the variable WEIGHT_SELF_DATE will contain the date in format YYYYMMDD and the variable WEIGHT_SELF_DATE_A will contain code 1.

If the day is missing, you have to code 15 for the day. The variable WEIGHT_SELF_DATE will contain the date in format YYYYMM15 and the variable WEIGHT_SELF_DATE_A will contain code 2.

If both day and month are missing, you have to code 0701 for month and day. The variable WEIGHT_SELF_DATE will contain the date in format YYYY0701 and the variable WEIGHT_SELF_DATE_A will contain code 3.

Variable names

Each variable will have a unique name in the database.

It is important that we all use the same variable names in the database. By applying the same variable names and value formats to variables:

- the misunderstandings and mistakes will be fewer,
- the programming will be more efficient,
- resource persons, such as database managers, programmers and statisticians, can be used more efficiently, with less time spent on deciphering

In the long run, the documentation and archiving of data will be much improved. Databases will be more valuable and useful for all research in the future when the information on the data is less dependent on the actual collectors.

The variable names should be as short as possible without losing its description of data. Max length is 18 digits. The variable names should, as far as possible, explain what data it is in the variables. For that reason has variable names with number not been used. For example: The names for the variables "Weight - self reported" and "Weight – measured" are WEIGHT_SELF and WEIGHT_ME. These names are more self explanation than names like WEIGHT01 and WEIGHT02

Data format

Variable name	Description	Type	Length	Values
BIRTH_DATE	Date of birth	Date	8	YYYYMMDD - Year, month, day YYYYMM15 - day unknown YYYY0701 - day and month unknown NULL/blank - Irrelevant/non-participant/non-response, Irrelevant/structural, Do not know, Unknown
BIRTH_DATE_ACCURACY	Accuracy code for Date of birth	Number	1	1 - Date is complete 2 - Day is missing, coded 15 3 - Day and month are missing, cocoded 0701 6 - Irrelevant
BIRTH_LENGTH_REG	Birth length - midwife/register information	Number	2	10-80 - Length in centimetres 98 - Do not know 99 - Unknown Blank/NULL - Irrelevant, non participant, non-response
BIRTH_LENGTH_SELF	Birth length – self reported	Number	2	10-80 - Length in centimetres 98 - Do not know 99 - Unknown Blank/NULL - Irrelevant, non participant, non-response
BIRTH_WEIGHT_REG	Birth weight – midwife/register information	Number	4	100-8000 - Weight in grams 9998 - Do not know 9999 - Unknown Blank/NULL - Irrelevant, non participant, non-response
BIRTH_WEIGHT_SELF	Birth weight – self reported	Number	4	100-8000 - Weight in grams 9998 - Do not know 9999 - Unknown Blank/NULL - Irrelevant, non participant, non-response

Variable name	Description	Type	Length	Values
CENTRE	Centre	Text	2	<p>Australia:</p> <p>Denmark: 01 – The Danish Twin Registry</p> <p>Finland: 01 – Dept. of Epidem. And Health promotion 02 – Dept. of Public Health 03 – Finnish Genome Centre</p> <p>Italy: 01 – Italian National Institute of Health</p> <p>Netherlands: 01 – Vrije University</p> <p>Norway: 01 – Norwegian Institute of Public Health</p> <p>Sweden: 01 – Karolinska Institutet, The Swedish Twin Registry 02 – Uppsala Mol Med 03 – Uppsala Rudbeck lab</p> <p>UK: 01 – National Health Service</p>

Variable name	Description	Type	Length	Values
CHECKSUM	Checksum A subset of the EUIDNUM*	Number	1	See section EUid number .
COHORT	Cohort	Text	2	<p>Denmark: 01 – The cohort born in 1870-1996</p> <p>Italy: 31 - Population based. Twins born in 1920-1940, both are alive in the 2002 and they are living in Rome city (F. Giubilei). 32 - Population based. Twins born in 1920-1940 both are alive in the 2002 and they are living in Latina county (S. Giampaoli). 11 - Population based. Twins born in 1985-1994 both are alive in the 2003 and they are living in Milan and Lecco county (M. Battaglia).</p> <p>Sweden: 01 – The cohort born in 1886-1925 02 – The cohort born in 1926-1958 03 – The cohort born in 1959-1990</p>
COUNTRY	Country code where the data is collected. NOT where the twin is born. A subset of the EUIDNUM*	Text	3	036 – Australia 578 – Norway 208 – Denmark 752 – Sweden 246 – Finland 826 – UK 380 – Italy 999 – Unknown 528 – Netherlands
DEATH_DATE	Date of death	Date	8	YYYYMMDD - Year, month, day YYYYMM15 - day unknown YYYY0701 - day and month unknown NULL/blank - Irrelevant/non-participant/non-response, Irrelevant/structural, DNK, Unknown

Variable name	Description	Type	Length	Values
DEATH_DATE_ACCURACY	Accuracy code for date of death	Number	1	1 - Date is complete 2 - Day is missing, coded 15 3 - Day and month are missing, coded 0701 6 - Irrelevant
EUIDNUM	The unique identification numbers.	Text	12	See EUid number
GENDER	Gender	Number	1	1 - Male 2 - Female
HEIGHT_ME	Height – measured	Number	3	10-300 - Height in centimetres 998 - Do not know 999 - Unknown Blank/NULL - Irrelevant, non participant, non-response
HEIGHT_ME_DATE	Measured height – date	Date	8	YYYYMMDD - Year, month, day YYYYMM15 - day unknown YYYY0701 - day and month unknown NULL/blank - Irrelevant/non-participant/non-response, Irrelevant/structural, DNK, Unknown
HEIGHT_ME_DATE_A	Accuracy code for Measured height – date	Number	1	1 - Date is complete 2 - Day is missing, coded 15 3 - Day and month are missing, coded 0701 6 - Irrelevant
HEIGHT_SELF	Height – self reported	Number	3	10-300 - Height in centimetres 998 - Do not know 999 - Unknown Blank/NULL - Irrelevant, non participant, non-response
HEIGHT_SELF_DATE	Self reported height – date	Date	8	YYYYMMDD - Year, month, day YYYYMM15 - day unknown YYYY0701 - day and month unknown NULL/blank - Irrelevant/non-participant/non-response, Irrelevant/structural, DNK, Unknown

Variable name	Description	Type	Length	Values
HEIGHT_SELF_DATE_A	Accuracy code for Self reported height – date	Number	1	1 - Date is complete 2 - Day is missing, coded 15 3 - Day and month are missing, coded 0701 6 - Irrelevant
IDNUM	Identification number A subset of the EUIDNUM*	Number	1	1 – Twin 1 2 - Twin 2 3 - Triplet 4 - Quadruplet The indicator for individuals being none twins will be: 0 - Non twins
RANDOMNUM	Randomized number A subset of the EUIDNUM*	Text	7	See section 5.6 Clarification of some variables. EUID number
VITALSTATUS	Vital status	Number	1	1 - Alive 2 - Disappeared 3 - Emigrated 4 - Died 9 - Unknown
WEIGHT_ME	Weight – measured	Number	3	1-800 - Weight in kilos 998 - Do not know 999 - Unknown Blank/NULL - Irrelevant, non participant, non-response
WEIGHT_ME_DATE	Measured weight – date	Date	8	YYYYMMDD - Year, month, day YYYYMM15 - day unknown YYYY0701 - day and month unknown NULL/blank - Irrelevant/non-participant/non-response, Irrelevant/structural, DNK, Unknown

Variable name	Description	Type	Length	Values
WEIGHT_ME_DATE_A	Accuracy code for Measured weight – date	Number	1	1 - Date is complete 2 - Day is missing, coded 15 3 - Day and month are missing, coded 0701 6 - Irrelevant
WEIGHT_SELF	Weight – self reported	Number	3	1-800 - Weight in kilos 998 - Do not know 999 - Unknown Blank/NULL - Irrelevant, non participant, non-response
WEIGHT_SELF_DATE	Self reported weight – date	Date	8	YYYYMMDD - Year, month, day YYYYMM15 - day unknown YYYY0701 - day and month unknown NULL/blank - Irrelevant/non-participant/non-response, Irrelevant/structural, DNK, Unknown
WEIGHT_SELF_DATE_A	Accuracy code for Self reported weight – date	Number	1	1 - Date is complete 2 - Day is missing, coded 15 3 - Day and month are missing, coded 0701 6 - Irrelevant
ZYGOSITY	Zygoty	Number	1	1 - Monozygotic 2 - Dizygotic same sex 3 - Dizygotic opposite sex 9 - Unknown
ZYGOSITY_ASS	Zygoty assessment method	Number	1	1 - Based on genetic information (DNA) 2 - Based on similarity questionnaire from both twins 3 - Based on similarity questionnaire from one twin 4 - Based on serological analyses 8 - No Information 9 - NA because opposite sex twin

Table 1 Meta Data

Appendix

Appendix A – Migrain

Appendix B – CHD

Appendix C - Stroke