



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

Version: 3.2.4

Ann Björklund
Jan-Eric Litton

Table of contents

1.	Background.....	3
2.	Introduction	3
3.	Updates.....	4
4.	Aims	5
5.	Data specification.....	6
5.1	Appendix	6
5.2	Missing data.....	6
5.3	Error codes for dates	6
5.4	Variable names.....	6
5.5	Data format	7
5.6	Clarification of some variables	13
6.	Transferring data.....	15
6.1	File format	15
6.2	File name	15
6.3	Help	15
6.4	Security.....	15
7.	Appendix	15
7.1	Appendix that are finished.....	15
Table 1 File Format		13

1. 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.

The aims of this data standard are:

- To facilitate the work in the planning, analysis and archiving stage of GenomEUtwin.
- To increase accuracy of data.
- To increase comprehensibility of data and thereby enable a smooth transfer of datasets between co-workers.

2. Introduction

The database core decided during the meeting 18th-19th of November 2002 that a prototype of a common twin database would be built.

At the beginning the database will consist of following phenotypes:

- EUid number
- Country code
- Center
- Cohort
- Randomized number
- Identification number
- Date of birth
- Gender
- Date of death
- Zygoty
- Birth weight
- Birth length
- Weight
- Height

Each centre will contribute with data of 100 twins.

In the beginning of the summer 2004 (preliminary) real data (phenotypes above and migraine) for all twins will be collected from each centre.

3. Updates

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.

Since version 3.2.2

Two new centres from Sweden are added in the variable CENTRE: 02 – Uppsala Mol Med and 03 – Uppsala Rudbeck lab.

Since version 3.2

Some of the variables in section 5.5 Data format are rearranged and the variable CHECKSUM is added.

Length and type is changed for the variable VERSION.

The example in section 5.5 is corrected.

Since version 3.1

The structure of the Euidnumber is new. See section [5.5 Clarification of some variables](#)

The max length of the variables is changed from 12 to 18. See section [5.4 File format](#). This results in that the variables have got new variable names.

The variables COUNTRY, CENTRE, and COHORT have changed data type to TEXT.

The country code for Australia and Finland has been corrected.

The variable BIRTH_PLACE has omitted.

The error code -97 – *Irrelevant structural* has been omitted in the variables BIRTHWEIGHT_SELF, BIRTHWEIGHT_REG, BIRTHLENGTH_SELF, BIRTHLENGTH_REG, WEIGHT_SELF, WEIGHT_ME, HEIGHT_SELF, HEIGHT_ME,

Since version 3.0

This document has got a new name, Data Format and Variable Standard for GenomEUtwin's Phenotype Database Prototype.

The GenomEUtwin number has change name to EUid number. The twinnumber, that is included in the EUid number, has changed format and name. Now it contain of 2 digits and the new name is Individual number.

The country code in the EUid is where the twin data is collected NOT where the twins are born.

Short explanations of the different types of missing data that can exist are given in section [5.2 Missing data](#).

An error code variable is added to all date variables. A proposal how to code incomplete dates are added in section [5.3 Error codes for dates](#)

There is a proposal for standard variable names included in the table in section [5.5](#).

Since version 2.4

Codes for form and version are added in section [5.4 File Format](#).

The eutwinnum is changed:

- The country code has 3 digits.
- The centre has 2 digits.
- The cohort has 2 digits.

There are new codes for all variables that contain dates.

The variable Vital Status has changed code for the answer unknown.

Variable names is included in the file specification, see chapter 5

There are two new codes (triplet, quadruplet) added for the variable twinnumber.

Code for "Have not been collected" has been added to the height and weight variables.

Chapter [6 File Transfer](#) has increased.

Since version 2.3

Codes for Birth length are added.

Since version 2.2

Codes for checksum changed to one digit.

Since version 2.1

Codes for country code, centre, cohort, pairnumber/familynumber and twinnumber are added.

Code 5 "Unknown" is added for Vital Status.

Since version 2.0

Codes for Vital status are added.

"Birth weight" is split up in two parts and the new parts are "Birth weight – self reported" and "Birth weight – midwife/register information".

"Weight" is split up in two parts and the new parts are "Weight – self reported" and "Weight – measured".

"Self reported weight – date" and "Measured weight – date" are added.

"Height" is split up in two parts and the new parts are "Height – self reported" and "Height – measured".

"Self reported height – date" and "Measured height – date" are added.

Since version 1.0

Birth weight is changed to four digits. And the codes for "Do not know" and "No answer" are changed. The new codes are: "Do not know" – 9998, "No answer" – 9999.

There are new codes for Zygoty.

4. Aims

The aim is to define the variables and the variable types of which the database and transfers files will consist of. This file will be sent from each centre to The Swedish Twin Registry, Ann Björklund, ann.bjorklund@meb.ki.se.

5. Data specification

5.1 Appendix

All phenotypes except the basic ones that are in this document will be in different Appendix. See chapter 7. Appendix

5.2 Missing data

In the data format in section 5.5 are different codes used for coding different reasons for missing data. These different types of missing data are 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.

Comments for extraction:

When the database is used for extraction for use, e.g. in MX or STATA, we will be able to provide missing data in a specified format. For example, if so desired all missing data, regardless of reason, can be output as -9999.

5.3 Error 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 error code variable added after each date variable.

For example:

The variable WEIGHT_SELF_DATE (Self reported weight – date) has an error variable named WEIGHT_SELF_DATE_E.

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_E 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_E 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_E will contain code 3.

5.4 Variable names

Each variable will have a unique name in the database and in the transfer files.

It is important that we all use the same variable names in the database and transfer files. 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. Therefore is a column with variable names added in the table in section [5.5 Data Format](#).

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

5.5 Data format

Variable name	Description	Type	Length	Values
FORM	Form number	Number	2	See section 5.5 Clarification of some variables .
VERSION	Version of the File Specification	Text	5	3.2.2
COUNTRY	Country code where the data is collected. NOT where the twin is born. A subset of the EUIDNUM*	Text	3	036 – Australia 208 – Denmark 246 – Finland 380 – Italy 528 – Netherlands 578 – Norway 752 – Sweden 826 – UK 999 – Unknown
RANDOMNUM	Randomized number A subset of the EUIDNUM*	Text	7	See section 5.6 Clarification of some variables .

Variable name	Description	Type	Length	Values
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
CHECKSUM	Checksum A subset of the EUIDNUM*	Number	1	See section 5.6 Clarification of some variables.
CENTRE	Centre	Text	2	Australia: Denmark: 01 – The Danish Twin Registry Finland: 01 – Dept. of Epidem. And Health promotion 02 – Dept. of Public Health 03 – Finnish Genome Centre Italy: 01 – Italian National Institute of Health Netherlands: 01 – Vrije University Norway: 01 – Norwegian Institute of Public Health Sweden: 01 – Karolinska Institutet, The Swedish Twin Registry 02 – Uppsala Mol Med 03 – Uppsala Rudbeck lab UK: 01 – National Health Service

Variable name	Description	Type	Length	Values
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>
BIRTHDATE	Date of birth	Date	8	<p>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</p>
BIRTHDATE_ERROR	Error code for Date of birth	Number	1	<p>1 - Date is complete 2 - Day is missing, coded 15 3 - Day and month are missing, coded 0701 6 - Irrelevant/non-participant/non-response 7 - Irrelevant, structural 8 - Do not know 9 - Unknown</p>
GENDER	Gender	Number	1	<p>1 - Male 2 - Female</p>
VITALSTATU	Vital status	Number	1	<p>1 - Alive</p>

Variable name	Description	Type	Length	Values
S				2 - Disappeared 3 - Emigrated 4 - Died 9 - Unknown
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
DEATH_DATE_ERROR	Error 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/non-participant/non-response 7 - Irrelevant, structural 8 - Do not know 9 - Unknown
ZYGOSITY	Zygoty	Number	1	1 - Monozygoty 2 - Dizygoty same sex 3 - Dizygoty 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
BIRTHWEIGHT_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
BIRTHWEIGHT_REG	Birth weight – midwife/register information	Number	4	100-8000 - Weight in grams 9998 - Do not know 9999 - Unknown Blank/NULL - Irrelevant,

Variable name	Description	Type	Length	Values
				non participant, non-response
BIRTHLENGT H_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
BIRTHLENGT H_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
WEIGHT_SEL F	Weight – self reported	Number	3	1-800 - Weight in kilos 998 - Do not know 999 - Unknown Blank/NULL - Irrelevant, non participant, non-response
WEIGHT_SEL F_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_SEL F_DATE_E	Error 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/non- - participant/non-response 7 - Irrelevant, structural 8 - Do not know 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

Variable name	Description	Type	Length	Values
				YYYYMM15 - day unknown YYYY0701 - day and month unknown NULL/blank - Irrelevant/non-participant/non-response, Irrelevant/structural, DNK, Unknown
WEIGHT_ME_DATE_E	Error 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/non-participant/non-response 7 - Irrelevant, structural 8 - Do not know 9 - Unknown
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
HEIGHT_SELF_DATE_E	Error 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/non-participant/non-response 7 - Irrelevant, structural 8 - Do not know 9 - Unknown
HEIGHT_ME	Height – measured	Number	3	10-300 - Height in centimetres 998 - Do not know 999 - Unknown Blank/NULL - Irrelevant, non participant, non-response

Variable name	Description	Type	Length	Values
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_E	Error 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/non- - participant/non-response 7 - Irrelevant, structural 8 - Do not know 9 - Unknown

Table 1 File Format

5.6 Clarification of some variables

Form (FORM)

The number of this data form is 3. Data form 1 and 2 is described in the document for sample collection, written by Anne Leinonen.

Version (VERSION)

The number of the version of this document.

EUid number (EUIDNUM)

The final EUidnumber consists of four parts:

- Country code 3 digits – ISO 3166
- Randomized number 7 digits
- Identification number 1 digit
- Check sum 1 digit

NOTE: The EUidnum is not delivered as an own variable, it is only delivered as the sub variables (COUNTRY, RANDOMNUM, IDNUM, CHECKSUM) separately.

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 EUidnumber, 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 EUidnumber needs indicator weather or not this is a twin. To obtain this information and still have unique numbers the EUidnumber 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

6. Transferring data

6.1 File format

Data are prepared in ASCII comma-delimited format using semicolon (;) as a delimiter with variable names in a first row. If a value contains text with spaces or semicolons (;) then all text is surrounded by double quotes (").

6.2 File name

To simplify tracing of data transfers between GenomEUtwins centres is recommended to name data files as follows:

GT_country_centre_cohort_YYYYMMDD_NN, where

GT	identification of GenomEUtwins project
Country	value of COUNTRY variable, as it appears in data file
Centre	value of CENTRE variable, as it appears in data file
Cohort	value of COHORT variable, as it appears in data file
YYYYMMDD	year, month and date when file was created
NN	sequential number of transfers

6.3 Help

The **GenomEUtwins Data Transfer Java application** can be used to analyse and prepare data files for sending them by e-mail. You can launch this application from here: <http://www.ktl.fi/morgam/apps/genomeutwin/> .

6.4 Security

To avoid that non authorized can get the twin information is it necessary that you email the file PGP encrypted. Please send the file to The Swedish Twin Registry, ann.bjorklund@meb.ki.se. The public key is attached in the same email as you get this document.

If you have any problem to encrypt the data, please contact Zygimantas Cepaitis, zygimantas.cepaitis@ktl.fi

7. Appendix

7.1 Appendix that are finished