Introduction

This document is the Data Dictionary for the data from the Airwave Health Monitoring Study database. It provides an overview of the conventions used for the export as well as annexes describing each data field.

This is a document-set in continuing development, and we shall add more datasets as and when we can.

We are interested in feedback from researchers on the approach taken generating these extracts. Conventions used have been based on the preferences of the first batch of researchers, and these may not be suitable for everyone. For example, our treatment of missing values may not be convenient for some.

Document Configuration

Subject	Data Dictionary
Version	2.0
Author	Andrew Heard, Database Manager
Last Saved	20-May-2022 14:49 by Heard, Andy H
Revision Number	42
Pages	4

Changes at Version 2.0

The data-tables have now been relocated into separate annexes; and the procedural background to the data capture process is now in Appendix A.

Data now includes follow-up results where available.

Within the data itself, there have been a small number of corrections to individual values where further investigations revealed errors.

The Scottish cancer registry reported to us in January 2018 that the reported “treatment date” field was and is “incidence date”.

Data Types

Fundamentally, there are three types of data in this extract: character strings, numbers and dates. Strings and numbers have sub-types that are restricted versions of the super-type. These are all explained below.

Character Strings

String fields contain variable length character data from the ASCII code-set (letters, digits, punctuation and so on). There are three subtypes.

VARCHAR2 (length)

Variable length character fields not exceeding length (when stated), which will in any case be less than 1000. Values have been coerced to uppercase in this extract.

YESNO

Answers to questions that have a yes / no response. Literal values will be either YES or NO.

YESANY

These 3-character strings report YES if any one of a set of values is YES. This is useful when, for example, a participant has made several clinic visits and provided a urine sample on just one occasion. No other value is allowed.

Numbers

These are positive or zero floating point values with up to 38 digits of precision. Numbers are represented to the maximum number of places that they are reported to us. Rounding is performed on the derived values (means etc.), the number of decimals being one more than the least number of decimals in the constituent data types. For example, height is reported to one decimal place, and the mean is rounded to two. The subtypes are:

- NUMBER (precision): Integers with up to *precision* digits.
- NUMBER (precision, scale): Floating point numbers with up to *precision* digits and *scale* decimals. For example, the range of NUMBER (5, 3) is 0 ≤ value ≤ 99.999;

Dates

These fields store a date and time with a possible internal resolution of one second. They are represented in the extract in the format DDMONYYYY with an optional time component which is HH:MI:SS (24-hour clock).

The mean of derived values is computed using the unrounded values held internally. The mean is then reported with rounding as above.

Representation of Missing Values (Contingency Codes)

Researchers asked us, wherever possible, to differentiate the reasons why some values are missing. This section explains our solution. First, some definitions:

- A **literal** is any value that is not missing and which can be used for its intended purpose. So, a set of positive values for weight will be literals.
- A **contingency-code** is a placeholder for a value that is missing. Because we wish to maintain the convention that a column defined to be numeric, say, will contain only numeric data, we have defined a set of numeric values to represent the reasons for a value being missing. Users should ensure that they do not, therefore, average all weight values including the contingency codes (which are actually negative numbers).

We would be interested in hearing the opinion of the researcher community as to whether this is a good approach, or whether it makes analysis unnecessarily complex. An alternative, for example, would be to present a shadow file for each extract that contains only the contingency codes. The main extract would then include missing values as nulls.

Types of Contingency Code

Including the codes used in the surveys and questionnaires extracts[1], the different types of contingency are below.

Ex Protocol

A question was not in use in the version of the protocol in force at the time the participant was screened.

Not Applicable

A question was not asked because it would never be meaningful for the current participant. For example, we would not ask a male participant if they were pregnant, and any response would be discarded.

Not Collected

Not Collected is reported when a value is missing for reasons that are explained by other data held about the participant. However, unlike Not Applicable, we would have reported those values had they been present. For example, the protocol states that only participants reporting themselves as diabetic should provide a standing blood pressure measurement; but if one had been taken for a non-diabetic, it would be shown.

Not Found

There is no record of a result having been received and no clear explanation in the dataset to explain why. For example, data may have been lost due to technical or procedural failure.

Unusable

One or more responses to this question are present but they are all deemed in some way unreliable or otherwise faulty. For example, when a reported value is biologically impossible (see: Allowable Range of Values) without explanation.

Value Conflict

Two or more plausible but conflicting values were obtained for this question, and the usual rules for choosing between them have failed.

Redacted

These are fields whose values have been removed from the extract, usually to ensure the anonymity of participants.

Missing Section

For survey / questionnaire data, this means that the whole section of a survey or questionnaire was missing – probably because of a technical problem.

Optional

For survey / questionnaire data, this means that the question was stated to be optional and the participant chose not to answer.

[1] "Validation Rules & Statuses" – documentation of the Study's surveys and questionnaires.

Representation of Contingency Codes

There is one user-configurable value for each contingency code for each of the three fundamental data-types. We validate that no exported literal value is also a contingency code. In the current extract, the values are:

Contingency	Numbers	Strings	Dates
Ex Protocol	-1	_1_	31-Dec-1610
Not Collected	-2	_2_	31-Dec-1620
Not Found	-3	_3_	31-Dec-1630
Unusable	-4	_4_	31-Dec-1640
Value Conflict	-5	_5_	31-Dec-1650
Not Applicable	-6	_6_	31-Dec-1660
Redacted	-7	_7_	31-Dec-1670
Missing Section	-8	_8_	31-Dec-1680
Optional	-9	_9_	31-Dec-1690

Table 1: Contingency Codes

Exported Fields

The fields exported have been divided into functional groups and each is documented in its own annex. At present, these annexes have been published.

- **Annex A** - Screening Results: these are a set of results from the screening protocol, both the anthropomorphic data obtained at the clinic and those resulting from the initial laboratory assay of blood samples.
- **Annex B** – Employments: this is an extract of summary employment information that provide demographic information on participants subject to the need to ensure anonymity.
- **Annex C** – Electrocardiograms (ECGs): summary, group codes and Minnesota codes for the ECGs taken during the Study.