Introduction

This is a technical overview for researchers of the Airwave Health Monitoring Study's research datasets. It provides background on data collection and management methods as well as conventions applicable to each of the exported datasets. Detailed explanations of each dataset are in the extract-specific annex.

The next section addresses information governance issues likely to be of direct interest to researchers. Following that, we explain how participants and their data are identified and linked via pseudonymised variables. Then an overview of the most important quality assurance and data correction techniques we use in collecting, cleaning and validating data prior to export. Finally, we explain general concepts used in creating extracts and how files and records are structured.

We are interested in feedback from researchers on the approach taken generating these extracts, which are mostly based on the preferences of the first generation of researchers.

Version Configuration

The current version of this document is available online along with each of the annexes. This copy is revision 42, last saved on 27-May-2025.

Information Governance

The Airwave Study operates within an information governance framework that is explained in more detail in our data management plan. The following sections highlight certain topics likely to be of direct relevance to researchers using the datasets we produce. More information is available in the governance section of our website.

Pseudonymisation and Personal Data

All datasets have been pseudonymised before being made available to researchers. This means that directly identifying variables such as name and date-of-birth have been removed. Records are keyed using one of a small quantity of pseudonymous identifiers that appear across datasets, and which allow researchers to link across datasets. We describe these pseudonymised variables in more detail below.

From the perspective of a researcher, research datasets are functionally anonymous. However, they are still "personal data" under data protection law because a mapping exists to their identities. It is irrelevant that the researcher has no access to that mapping.

Disclosure Control Policy

Users of datasets are legally bound not to attempt to identify individual members of the cohort. Our disclosure control policy aims to minimise the risk of accidental re-identification by aggregating or repressing variables whose cardinality is five or fewer. We also look for groups of variables that, although passing this five-or-fewer rule on an individual basis, could potentially be used in combination with public domain information to identify someone. For example, we merge rank for the most senior officers because of its vulnerability when joined with gender, ethnicity, location and data collection date.

Please notify the Airwave data management team if you discover some combination of variables that an informed browser could reasonably use as a means of identification.

Withdrawals

Participants who withdrew fully from the cohort before the advent of the Data Protection Act 2018 – the GDPR update - have been removed from the cohort and all research datasets.

Those who have withdrawn consent to share their data with external research platforms such as DPUK and UK LLC are excluded from datasets. Updates will be issued from time-to-time.

Participants cannot remove themselves from research datasets used at Imperial College that support the principal aims of the Airwave Study. Their records are, however, anonymised rather than pseudonymised, and no further data updates will occur.

Identifying Participants and Research Engagements

This section provides background information on how study records are identified and linked across datasets.

Participants

A “participant” is a consenting individual who has joined the cohort either by completing an enrolment questionnaire or volunteering for a health screen. Participants are identified pseudonymously using a 7-digit value whose variable name is part_id.

Unique participants were derived from study data using basic demographic data such as name, address, date-of-birth and employment identifiers. Linkage was verified by NHS list cleaning services.

Enrolment Questionnaires

Enrolment questionnaires were paper documents containing a set of questions useful for the research, plus a consent form. Anyone who completed the questionnaire and signed the consent form is a member of the cohort. We scanned the documents and loaded transcribed copies of the relevant results into our database. Questionnaires are identified by an integer labelled qnr_id.

Clinic Barcode

Officers and staff who completed an enrolment questionnaire were offered a health-screen. These were conducted at locations convenient to the participant and were run by staff retained by Imperial College.

Participant consent to join the cohort was required by those taking and benefitting from the health screen. Although completing an enrolment questionnaire prior to the screen was optional, we strongly encouraged everyone to do so or take an equivalent tablet-based survey.

Every clinic visit is identified by its own a 5-digit integer which has come to be known and labelled as barcode. It is used to identify most clinic data, assays of biological samples and clinic-derived surveys.

At baseline, the physical cryovials holding biological samples were identified by barcode, shown on the tube in human-readable form and as an actual barcode. By the start of follow-up in November 2015, cryovials were being identified using a labware-specific identifier.

Extracts keyed on barcode will usually be cross-referenced within the extract to the part_id of the participant.

Validation of Collected Data

Producing a reliable research extract is a three-stage process.

- First, we setup the environment where data can be collected. A clinic, questionnaire or laboratory are all examples.
- Data collected during the operation of these environments are transferred to the central study data management team who consolidate them into a pre-link repository where they are assessed for accuracy. Once any corrections and exclusions have been completed, data are imported into the main cohort database where they are linked with existing datasets.
- Finally, we prepare consolidated extracts of data, grouping variables according to their likely usage by research teams. These are made available to researchers on Trusted Research Environments (TREs).

The remainder of this section briefly reviews the key validation processes we employ to ensure that data is as accurate and complete as it feasibly can be. Readers already confident in the quality of our data can skip these sections.

Validation at Environment Level

Each environment has its standard operating protocol (SOP) designed to ensure that data is accurately collected and recorded. For example, the nurses working in the clinics are all trained using a detailed SOP that describes the minutia of each data collection exercise and how to record the results.

Apart from the enrolment questionnaire, which was a paper document, all data have been collected on some form of electronic form such as a website or app. This allows for automated checking of data during the collection process, meaning that answers can be challenged and errors corrected at the point of collection.

Our methods for data validation were developed and enhanced as the study went on and can be categorized into the following stages.

- **Authentication** is designed to verify the identity of the person who is submitting the data, whether by proxy or in-person. Mistakenly linking new data with the wrong participant is a serious mistake if not corrected. Participants authenticating themselves to an online questionnaire will identify via a self-verifying code. They will authenticate using a password or by providing corroborating data such as date-of-birth.
- **Domain Checking**: entered data is generally keyed as text fields which are then converted into dates and numbers. We check that values offered are represented legally, so:
    - numbers are made up from digits, with or without a (single) decimal point.
    - dates are formatted legally.
    - character data is length-limited and, where necessary, chosen from a pre-defined set of values.
- **Sanity Checking**: values entered are checked for being reasonable in the circumstances. For example, questions relating to the age at which a medical diagnosis was made should be consistent with date of birth. A body weight of 928 Kg would be queried, allowing the respondent to add the decimal point before submitting.

- **Completeness**: some fields are regarded as mandatory and require a response to proceed.
- **Repeat Measurements**: measurements such as height and weight are made two or three times to ensure consistency and repeatability. Internal variation within each set of measurements must lie within plausible limits to be accepted.

Laboratory analysers have their own sophisticated internal checking algorithms that validate accuracy of the assay. We can provide technical details of the analysers used throughout the study to any interested party.

Validation by the Data Management Team

Some of the checking conducted at this stage is the same as the previous stage (range checking, missing values etc.) but we also take a view on borderline cases in the context of all the results collected that day by the nurse or using that equipment. Flagged cases may be examined by clinical and data team members, possibly in consultation with the nurse. For example, we reassured ourselves that an exceptionally low set of height measurements were in fact correct when the nurse confirmed a case of dwarfism not otherwise noted.

Inevitably, we occasionally receive duplicate and alternative data that differ from the earlier submission. We have policies for these cases to ensure consistency. For example, a borderline out-of-range laboratory assay might be repeated if, in the opinion of the laboratory manager, there was sufficient reason. If we are then advised by the laboratory to use the second result in place of the first, we would do so. If the second result was consistent with the original, we would use the first one.

An alternative explanation for a duplicate result is a faulty barcode. This triggers a local investigation and data correction when identified.

The data management team occasionally identify subtle types of error that occur from time to time, and which weren't envisaged during design of the data collection tool. For example, in the example below, the nurse has mixed up rows and columns when entering the results of waist and hip measurements. When we noticed this occurring more than once, we redesigned the data collection tool to challenge this class of error at the point of entry.

	Intended			Entered	
Entry	Waist	Hip		Waist	Hip
1	90.5	100.4		90.5	90.1
2	90.1	100.1		100.4	100.1

Table 1: Example of Data Entry Error

During baseline we computed CUSUM scores of numerical data for the baseline datasets to highlight issues with laboratory assays and / or nurse technique. This is a simple but powerful method of identifying subtle changes resulting from systematic drift over time where changes of individual values are insufficiently large to be flagged as problematic. Issues raised in this way were used in quality control discussions with data providers and nurses. However, it is a feature of the CUSUM method for false signals to arise, and any problems need to be explainable by reference to the underlying method and the results of any control samples assayed.

Linkage Errors

Some kinds of errors are noticeable only when a day's results are considered together. This might happen, for example, when a clinic reports a barcode value that was assigned to a different clinic. Another example is a laboratory assay using a barcode for which no corresponding blood collection took place.

On investigation, these kinds of error are generally easy to correct by reference to supporting records such as paper documentation submitted by the clinic.

Auditing Changes

Overall, relatively few corrections have been needed. Where they have taken place, they are recorded in an audit trail identifying who made the change, the old and new values, when the change was made and descriptive text describing the circumstances.

Clinical Signoff and Participant Feedback

A popular feature of the clinic protocol is full disclosure to the participant of all medically useful results collected during the clinic and assayed by the subsequent laboratory analyses. As well as highlighting potential data errors, this also allows us to advise participants to seek medical consultation where appropriate.

Supported by a report on measurements and assays that lay outside the reference-range (two standard deviations from population mean), the study's clinical lead examined all the results for internal consistency in a medical context. As a result, we occasionally asked our laboratory to repeat a set of measurements using a stored sample; and faulty results that may alarm the participant were suppressed.

Creating an Extract

The final stage is to create an extract of consolidated, pseudonymised research results. This process builds on all the data cleansing work conducted at the clinic or laboratory, and the work done in the following days and weeks by the data management team.

The first stage of extract generation is to group variables together into sets that would seem to make sense in a research context. Of course, there are no uniquely correct solutions to this, and we are happy to receive feedback on our choices.

Where a variable has been collected over many years and there has been change in the collection method, we may need to provide supporting information as its own variable. In one case – HbA1c – we have provided a completely new variable for the assayed value, as the "old" and "new" methods are too different.

Data Conventions

From June 2025, datasets will, by default, be encoded in UTF-8 format using the windows-style end-of-line convention (CRLF). They are structured as a header record followed by multiple body records. The header comprises variable labels reported in upper-case. Each variable within a record is separated by a comma, and text fields enclosed by double-quotes. Prior to June 2025, most datasets were tab-separated, but the preference seems to be for CSV over TSV.

The following is a brief description of how variable types are coded.

Strings

String fields contain variable length character data, generally reported in upper-case. Values should not exceed the length stated in the relevant annex. There are two subtypes:

- **YESNO:** Answers to questions that are Boolean in nature. Values will be “YES” or “NO”.
- **YESANY:** 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 sample on at least one occasion.

Numbers

These are floating point values reported using the same number of significant digits that we collect. Rounding is performed on derived values, the quantity 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.

Dates

These variables provide a date and optional time whose maximum internal resolution is one second. Their format is “DDMONYYYY HH:MI:SS” (24-hour clock).

Representation of Missing Values

In most large datasets there are variables within records for which there is no usable value. When creating an extract, one option is to report null – the empty string. In cases where there is one reason only for a value being missing, we will often do just this and set out the reason in the annex.

When more than one possible reason exists for a missing value, and we have a basis for discriminating between reasons, we instead report a sentinel value. The following is a table of all sentinel values present in the Airwave dataset.

Reason	Description	Value
Ex Protocol	Variable was not required by the protocol in use at the time.	_1_
Not Applicable	A question was not asked because it would never be meaningful for the current participant. For example, we would not ask a staff member questions that were only relevant to police officers.	_2_
Not Collected	A value is missing for reasons 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 need provide a standing blood pressure measurement; but if one had been taken for a non-diabetic, it would nevertheless be shown.	_3_
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.	_4_
Unusable	One or more responses to this question are present but they are all deemed in some way unreliable or otherwise faulty. For example, assay values that are biologically impossible.	_5_
Conflict	Two or more plausible but conflicting values were obtained and the usual rules for choosing between fail.	_6_

Reason	Description	Value
Redacted	Values have been removed to ensure anonymity.	_7_
Missing Section	A whole section of a survey or questionnaire was missing – probably because of a technical problem.	_8_
Optional	A survey question was stated to be optional, and the participant chose not to answer.	_9_

Table 2: Sentinel Values

Obviously, for any one variable, only a subset may apply. Consult the annex for any variable-specific guidance.