Introduction

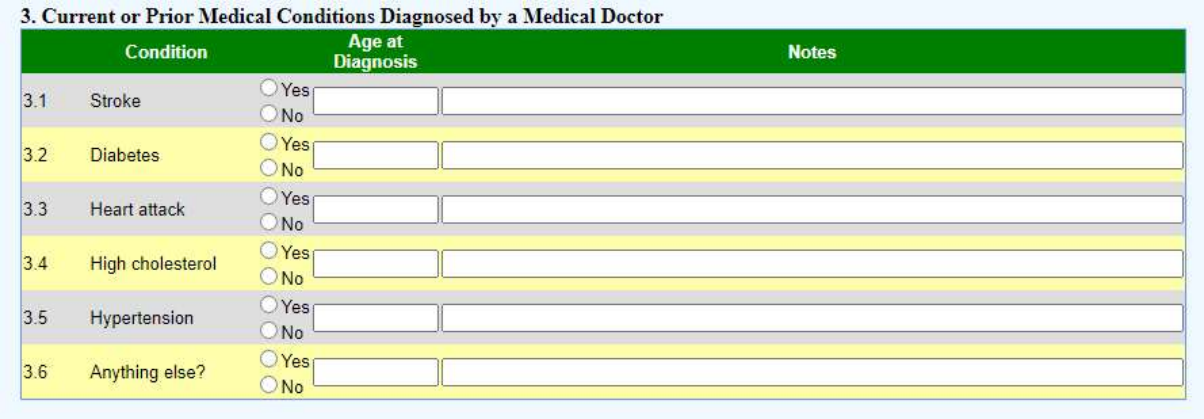
This Annex provides an explanation of the data-fields in the Airwave Study self-reported diagnoses export. The Background section contains important context that may be relevant to your analysis.

Title	Data Dictionary, Annex N
Subject	Metadata for Airwave Study self-reported diagnoses export
Version	1.0
Author	Heard, Andy H, Database Manager at Imperial College London. a.heard@imperial.ac.uk
Published	28/06/2024

Background

These data were collected to establish participants' basic medical histories. At the time of the rollout, relatively few datasets were available from NHS, so these data important were important in establishing pre-existing conditions.

Questions were put to participants at screening clinics by a nurse who recorded the answers on an electronic form. For the majority of the rollout, the form looked like Figure 1.



3. Current or Prior Medical Conditions Diagnosed by a Medical Doctor

	Condition		Age at Diagnosis	Notes
3.1	Stroke	○ Yes ○ No		
3.2	Diabetes	○ Yes ○ No		
3.3	Heart attack	○ Yes ○ No		
3.4	High cholesterol	○ Yes ○ No		
3.5	Hypertension	○ Yes ○ No		
3.6	Anything else?	○ Yes ○ No		

Figure 1: Nurse Form Capturing Diagnoses Data

These questions were intended to be specific to the participant. However, in the pilot phase of the study (2004 – 2006), the question was asked within the context of "… current or previous illnesses or conditions that affect you **or your family**" (my emphasis). This was not a great protocol, as we must now rely on the unstructured Notes to differentiate between self and family. We can't share the notes here for data privacy reasons, but we have used them to determine whether it is the participant or their family that is being referred to. Specifically, if the note contains any of the following keywords, it's determined to be a family member; otherwise, it relates to the participant.

father|brother|mother|sister|uncle|aunt|family|parent|cousin|grand

Conditions

The five conditions we asked about are shown in Figure 1 but there are some caveats that may be relevant to your analysis.

- **Currency of Diagnosis:** Just because a diagnosis is reported as positive, it doesn't necessarily mean that the participant had the condition at the time of the screen. It's clear from the comments that some diabetes cases, also high-cholesterol and hypertension, were transient and that the participant was free from the condition when asked. We suggest you refer to our screening dataset, which contains data that allow you to determine their current state of health for these three conditions.
- **Diabetes Types**: From 2004 – 2011 we asked simply about "diabetes". From February 2011 we started to ask specifically about type-1 and type-2.
- **Heart Attack:** In comments where participants answered "yes" to having suffered a "heart attack", it's clear that the term has been interpreted to mean many kinds of cardiac-related conditions. These include congenital defects and benign murmurs, not just myocardial infarctions. If you need more detailed data on cardiac health, you should refer to the electrocardiogram dataset.
- **Anything Else:** In the pilot (2004 – 2006), this option did not exist. It was then added so that participants could share their wider concerns, perhaps in case of contraindications to the clinic tests. Because of the difficulties in parsing and sharing free-form text, we have not attempted to interpret these additional diagnoses.

Variables

Label	Data Type	Description
barcode	NUMBER (5)	Pseudonymous identifier for the screening visit.
part_id	NUMBER (7)	Pseudonymous identifier that is unique for each participant. Because the screening extract includes follow-up visits, some part_id values are associated with two or more barcodes.
subject_id	INTEGER	Anonymised identifier for the participant that will replace part_id in the next version of this export.
gender	STRING	The most recent gender of the participant. This may differ from the gender assigned at birth.
condition	STRING	The condition diagnosed from the list above.
is_diagnosed	YESNO	"Y": diagnosis of this condition was made. "N": not diagnosed with this condition.
relates_to	STRING	"SELF": diagnosis is specific to participant. "FAMILY" a family member's diagnosis only.
when_reported	DATE	When the data was collected – mostly the clinic appointment date. This is NOT the date of diagnosis.
age_when_reported	NUMBER	Participant age (years) at when_reported (computed).
age_when_diagnosed	NUMBER	Participant's age at the time of diagnosis. This can be empty if the participant didn't remember accurately, or if the value is clearly faulty (age_when_diagnosed > age_when_reported, for example).

Table 1: Self-Reported Diagnoses Fields

File Versions

This is revision 1 of the dataset and it was extracted in June 2024. It comes in two versions:

- **Full** includes all participants in the Study for whom we have current consent to process their data. It is available exclusively on the Enclave, Imperial College's Trusted Research Environment (TRE).
- **A filtered** excludes participants who have asked us to not share their data with external TREs. It is otherwise identical to the full version.

Summary information on each version is in Table 2.

Version	Records	Barcodes	Participants	Size (Bytes)	CRC32
Full	307,000	61,896	45,542	17,387,649	20985CDA
Filtered	305,938	61,695	45,379	17,327,508	9BE3D82C

Table 2: File Versions' External Characteristics