Introduction

This document is the Data Dictionary for the health-screening data from the Airwave Health Monitoring Study database. It provides an overview of the data collection process as well as the label-by-label explanation of each item. The Annex[1] should only be relevant to those wishing to alter or recreate the extract.

Document Configuration

Subject	Data Dictionary
Version	2.0
Author	Andrew Heard, Database Manager
Last Saved	17-Apr-2018 14:16 by Andrew H. Heard
Pages	32

Changes at Version 2.0

The following have changed between the 2013 and 2015 extracts. There are, in addition, some changes to a small number of values as the result of ongoing data cleaning work.

- The change in assay for glycosylated haemoglobin has resulted in three new labels and, for clarity, renaming of three existing labels.
- The "medication" section and all its labels have been removed to a separate document on "treatments". This is now a much richer dataset following interpretation of free-format input and mapping to the British National Formulary.
- A protocol value of **FOLLOW-UP** is included to allow reporting of barcodes collected during the follow-up phase of the Study that began in November 2015. A new value of **HYBRID** is used for barcodes that include values for more than one protocol, usually as a result of a rebleed (a participant began the screen when one protocol was in force, but a revised protocol had been adopted by the time they had a rebleed).
- The computation of sitting-height has changed following a review of the height of the stools used.
- Waist and hip values are now reported when the participant declares themselves to be pregnant; previously we reported not-applicable.
- A fault in some values of hours_since_eat_blood has been corrected. This potentially affects barcodes where the individual had a rebleed. The description of this label explains certain difficulties that still exist in the use of these values.
- The method of determining the contingency code for missing values has been tightened up so that it more reliably detects when a value is missing because it was not a part of the protocol at the time.

Hyperlinks

Where hyperlinks are provided, the user should carry out their own research to establish that the content is suitable for their purposes.

[1] Annex One: Database Objects Used in Extract

Overview of Data Collection Process

This section provides an overview of data-collection methods used during the Study. Its purpose is to provide assistance in interpreting the extract without overwhelming the reader with details of the process used to create it.

Data Capture and Feedback Systems

There have been two data management systems: the first was written as a prototype, the second as the long-term solution. We outline both below.

Pilot System

When the Study was in its Pilot phase (June 2004 until August 2006) we built a set of VBA macros around an Excel spreadsheet to clean, consolidate, link and generate feedback letters. When exceptions arose, we would make enquiries of the nurse or laboratory and make the appropriate corrections. With volumes of data being relatively light during the Pilot, we were able to carry out error handling on a case-by-case basis without becoming overwhelmed by the task.

Once the Pilot System was decommissioned, its data was migrated to its own space on Oracle. It is still available in read-only form, if required.

Oracle System

Although operating at the very limit of Excel at the time, we learned much about the processing requirement during the Pilot. We used this experience to write a more sophisticated feedback system for Oracle. The aim was to automate as much of the process as possible, ensuring consistency between participants and providing support for larger workloads.

The Oracle System resides on the Study's Private Network and provides both a secure repository for storing the Study's data and the means of loading, cleaning, linking and reporting it. There are many separate subsystems (health screening is one such), but they all work on the same data.

Before reaching the Oracle System, the various data were captured and processed by several other systems. This process is described in more detail in the Study's System Level Security Policy[1].

Data Capture Concepts

The following outlines the events occurring to a health-screen record. The focus is on the Oracle System, but much of the discussion applied equally to the Excel spreadsheet, albeit with different implementation details.

Participants

A Study "participant" is an identifiable and consenting individual who has either completed an enrolment questionnaire and / or volunteered for one or more health screens. They are identified by a 7-digit part_id label beginning with the digit 1.

Because a participant can enter the cohort by several independent routes, we link Study records by identifying individuals within the larger number of health screen and questionnaire records. This is done using their personal and employment details, which are validated against an NHS database.

Barcodes

Every health-screen is identified by its own "barcode", which is normally a 5-digit integer (the small number of 4-digit barcodes were created early in the Study to identify non-Police participants). Although one barcode can only ever identify a single member of the cohort, some members of the cohort may have >1 barcode. This happens when participants are invited back for a "repeat screen" (usually several months after the initial screen) or if they have simply invited themselves back for another screen.

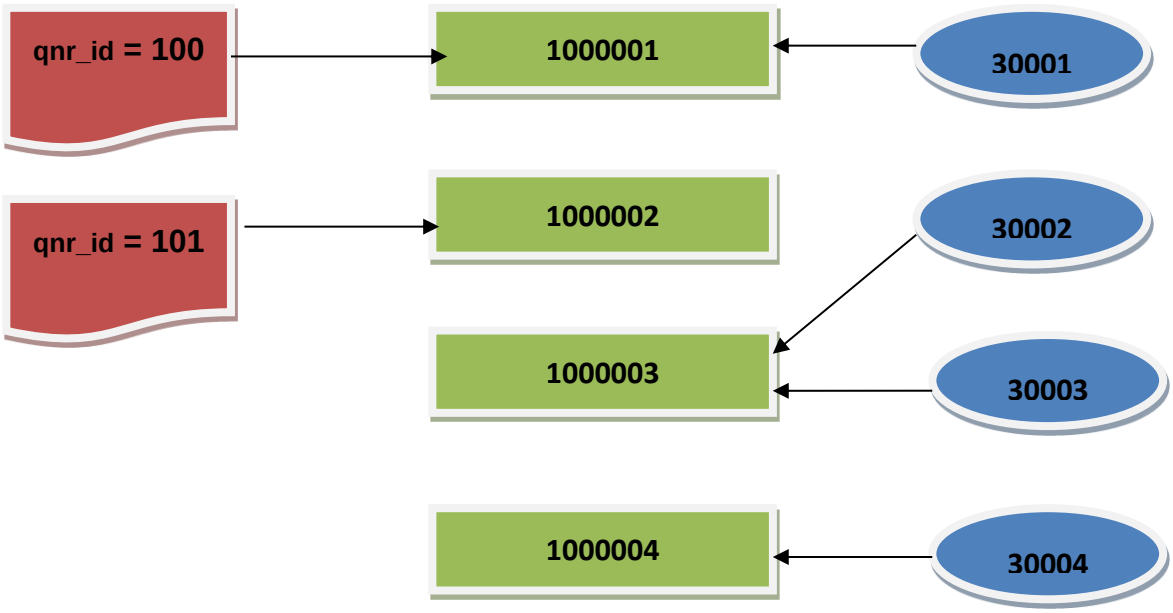

Figure 1: Links between Questionnaires, Participants & Screening Barcodes

In Figure 1, participant 1000001 has had a single health screen (barcode = 30001) that has been linked to an enrolment questionnaire (id = 100). Participant 1000002 is excluded from this extract because it has only a questionnaire record. Barcodes 30002 and 30003 are separate health-screens for the same individual. Participant and barcode 1000004 / 30004 is the more common case of a participant having a single screen.

Nurse Form

A volunteer makes an appointment at a clinic where they are assigned a "next off the top" barcode. A nurse carries out the screen and links all the information on the volunteer with their barcode. Most of the personal information - "nurse forms" - is submitted each day from a laptop to Imperial College where they are uploaded onto the Oracle System for processing.

Electrocardiogram (ECG)

Also collected at the clinic is an ECG reading, which is uploaded to Glasgow CARE for offline interpretation. ECGs are identified by barcode, summary personal information and timestamps. Batches of interpretations are returned to the Study once per month, where they are uploaded to the Oracle System and linked to the rest of the screening record.

Biological Samples

Most participants provide biological samples. These are physically labelled with the barcode and sent to the laboratory for analysis.

The results reported by the laboratory machines were originally printed off and retyped into a computer by the laboratory's own staff or the Study's administrators. An intranet-based HTML form was designed for this purpose.

Recognising the need to automate this process, we purchased a licence for imExpress™, a Windows-based software package from Data Innovations, LLC. This software collected data from the laboratory machines via their serial ports, which we converted into XML and integrated into Internet Explorer using ActiveX. This approach was implemented for the two main laboratory machines and was used until completion of the Pilot.

For the main Study, we upgraded our three analysers to modern equipment. After an initial period during which results from the machines could not be extracted electronically (back to retyping from hard copy!), an interface was based on spreadsheets and flat files was implemented. We receive and upload these results approximately once per week.

Feedback Results

The Oracle System attempts to consolidate each set of records by barcode to form a participant feedback letter. This will be returned to the volunteer and, optionally, their GP. It contains all the useful clinical results, signed-off by the Study's clinical lead.

Before the feedback letter is drafted, the Oracle System performs extensive validation of the dataset, checking for missing, faulty and duplicate values. To be available for feedback and closure, a barcode must have no outstanding errors, no duplicates and (subject to certain caveats explained in paragraphs below) no missing data.

The Study administrator investigates discrepancies and may amend results or fill in missing data according to an established process. Once a feedback has been completed, a participant identifier is assigned to the barcode, which is now considered closed. There is a process for reopening such barcodes, but it is unusual and is not relevant to this discussion.

Flagging to NHS Registers

Batches of participant data (name, sex, address, date-of-birth) are uploaded to the NHS once or twice a year for identity validation and flagging on the cancer registry. Feedback from this process assists in the linkage of barcodes to participants.

Files and File Types

The fundamental unit of data loading is a "file". These freestanding entities have properties of their own (barcode, filename, version, file-type, date-of-submission etc.) and a set of

records holding information on the barcode. Occasionally, the administrator adds single results manually onto the system, but even these are considered to belong to a special "system" file. There are typically 5 to 15 files per barcode.

Each file has a "type" property that is one of the following:

- **Laboratory:** Electronically recorded data from one of the analysers. Each Laboratory File is subdivided into one of four sub-types according to the type of machine that generated it (clinical chemistry; Eliza plate-reader; coagulation, or haematology).
- **Nurse Form:** The HTML document that captures the nurse's results on the laptop.
- **ECG:** interpreted ECG results from Glasgow CARE.
- **SLOG:** HTML form used at the laboratory and in the main office that allows users to hand-enter the results of laboratory analysis. SLOG forms were last used in February 2010.
- **SYS:** Derived values computed by the database or added directly by the administrator.

The file version is used to determine whether a value was "ex protocol" or not (see Interpretation of Contingency Codes on page 8).

Records

Each file loaded onto the Oracle System undergoes processing to detect errors and link to other records. Records contain either information about the file itself or measurements for the barcode.

Each record has two properties: a name field, cgi_name, which identifies the result being reported; and a value, cgi_value, which is the result itself. The set of permitted field names depends on the file-type and its internal version. The rules used to validate cgi_value depend on cgi_name and file type.

This is an extract of records from a nurse-form.

cgi_name	cgi_value	Comment
barcode	12345	Barcode of the participant whose results these are.
_revision	nscr_2.4.8	Identifies the type of file and its revision level, which is used by the Oracle System to validate the records.
bp_arm	RIGHT	A value stating that the right arm was used for the blood-pressure measurement.
bp_cuffsize	REGULAR	Data on the blood-pressure cuff used.
bp_sit_diastolic_1	93	The first diastolic blood pressure was 93.
h_timer_submit	06-MAR-2012 14:21:51	A timestamp generated by the laptop at the point of submission.

Table 1: Extract from a Nurse Form

Domain Checking Records

Every result enters the database as a string of characters, whatever its underlying data-type. We convert these characters into usable values according to the type we expect it to be. For example, we convert floating-point numbers into numeric data, ensure that integers have no decimals and that date fields are represented correctly. We also ensure that fields that are range-bound (restricted set of allowable values) hold a legitimate value. If validation fails, the record is flagged for investigation.

The data-types recognised by the system are:

- Floating-point number;
- Integer
- Date
- Date-time
- Free format character string
- Checkbox (a container for values that either exist or do not)
- Yes / No (a character string whose range is limited to yes or no responses)
- ECG Summary (a character string whose range is limited to the set of valid ECG interpretations returned by Glasgow CARE)

Sanity-Checking

The Oracle System flags records where the numeric or date value lies outside what we consider the "plausible range". This is designed to catch gross errors typical of manual data entry such as a missing decimal point. It then becomes an administrative task to check the value and either accept, amend or discard it. We have discarded values only in rare cases where the value is both "impossible" and it is not immediately clear what was intended.

The anthropomorphic measurements are measured and entered twice. In a small number of cases it is clear that the nurse has entered a value into the wrong box, as illustrated in Table 2. These values have been discarded. In later versions of the nurse-form we introduced warning systems that alerted the nurse to certain kinds of data entry error, and this does appear to have reduced their rate of incidence.

	What was Intended		What was 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 2: Example of Data Entry Errors

A later level of sanity checking occurs at the last step of feedback production. We produce an exception report that contains all values that lie outside the 95% reference range of clinically normal results. The Study's clinical lead then verifies, and may query, the feedback as a whole before it is mailed.

CUSUM Analysis

For most numeric results loaded since July 2006, we have computed a CUSUM score. Scores are computed for males and females separately, broken out by individual nurse where appropriate. The purpose is to identify systematic measurement errors and correct them.

Duplicates

A duplicate occurs when we have two measured values purporting to be for the same thing. Here are same example reasons (benign and otherwise):

- Participant names appear on both the ECG machine and the nurse form. Misspellings do sometimes appear in one or the other (usually the former).
- Nurses occasionally submit several forms for one barcode, marking one as a "trump" (corrections) file. We prefer the contents of any trump files to non-trump (the default).
- The laboratory may rerun a test on a sample if a result was unusual. We may then get two results, and we determine the one to report based on instructions from the laboratory.
- Some barcodes have multiple venepuncture events (rebleeds) because participants are unable to provide a complete sample at the first attempt. This may result in several blood results, one from each clinic visit.
- A file is incorrectly labelled with the wrong barcode. This is rare, as we use barcode scanners and other data verification techniques; however, it does still occasionally occur. We notice because one barcode has two sets of data for the same measurement whilst another has none. Wrong barcodes are also detected when files arrive on unlikely looking dates; for example, a laboratory file being time-stamped before the ECG was recorded is automatically flagged for investigation.

The Oracle System resolves some duplicates according to the rules that have been built into it. For example, we refer to any enrolment questionnaire to resolve ambiguity in the spelling of a surname. We also understand simple forename abbreviations (e.g. a forename such as "John" might be entered as "J." on the older ECG machines that laboured under a restricted name fields). We also use "fuzzy matching" algorithms such as SOUNDEX[2] to handle misspellings.

This process has been gradually refined over the years, but most of the development was completed within the first two years after the Oracle System's introduction. Anything that cannot be resolved automatically is investigated and resolved by an administrator or Database Manager.

Changes made to incoming data results in an audit trail identifying who made the change, the old and new values, when the change was submitted and often some descriptive text. When the Oracle System resolves duplicates automatically, it codes the reason for its determination onto the "ignored" record.

Missing Values

When data are missing, the administrator is invited to locate and upload the missing results. This usually involves discussions with the laboratory, nurse, Database Manager and Glasgow CARE.

If results are determined to be legitimately missing or irretrievably lost, the system can be instructed to issue the incomplete feedback. The system also applies certain rules itself. For example, a participant with an incomplete or failed venepuncture who agrees to return for a rebleed has 120 days from the date of the screen to do so. After this point, we waive the requirement that the laboratory results need to be complete, and draft the feedback anyway. If the participant does then subsequently have their rebleed, we will issue a revised feedback.

Some data can never be absent when generating a feedback. These are participant name, address, force-name, date-of-birth, sex, screening start-time, and consent.

Content of Extract

First, we differentiate between literal and Contingency Values (codes); then we explain about the data-types appearing in the extract.

Interpretation of Contingency Codes

Values that can be interpreted at face value (e.g. numbers that can be used to compute means, standard-deviations and so on) are called “literals”. Not every label for every barcode has a usable literal value. When one is not available, we report a “contingency code” that explains its absence.

There are five different types of contingency.

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 never ask a male participant if they were pregnant. Any response recorded to such a question would be discarded.

Not Collected

Not Collected is reported when a value is missing for reasons that are explained by others data held about the participant, but 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 provide a standing blood pressure measurement.

Not Found

There is no record of a result having been received and no clear explanation in the dataset to explain why. Reasons include:

- Data being lost at the clinic. Nurses have occasionally failed to save the final version of the form containing the participant's data, despite our many efforts to prevent this happening. Values that are available were obtained from one of the "interim" versions of the form, and others will be Not Found.
- A laboratory result is absent because there proved to be insufficient blood of a high enough quality to carry out the analysis.

Unusable

One or more responses to this question are present but they are all deemed in some way unreliable or otherwise faulty. For example, during signoff of a feedback letter it was determined that a value was highly implausible or likely to mislead. We might have attempted to ascertain a more likely looking result, but any such attempt failed.

The range within which a value must always lie is listed in Annex Two: Allowable Range of Values.

Value Conflict

Two or more plausible but conflicting values were obtained for this question, and the usual rules for choosing between them have failed. When the extract has been completed in its final version, this contingency will exist only for certain commentary fields (the nurse's final notes on completion of a clinic, for example).

Redacted

These are fields whose values have been removed from the extract. It is currently used only for free-format data whose content we cannot guarantee will be free of personally identifiable information.

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

Table 3: Contingency Codes

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). The maximum length of a string field is, in principal, 1000 characters; but at present, no field exceeds 500. There are three subtypes.

VARCHAR2 (length)

Variable length character fields not exceeding *length*. Values have been coerced to uppercase for the purposes of this extract.

YESNO

These fields code for 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. The values in the database may be {YES, NO, NO} and the extract will report YES. We report NO when all values are NO.

Numbers

These are floating point values in the range $0 \leq \text{value} \leq 9.999\times10^{125}$ with up to 38 digits of precision. The subtypes are below.

NUMBER (precision)

These are integers in the range 0 ≤ literal ≤ 10precision-1.

NUMBER (precision, scale)

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

Rounding of Floating Point Values

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 usually 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 fields store a date and time with an internal resolution of one second. They are represented externally in the form DDMONYYYY with an optional time component which is HH24:MI:SS.

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

Tables of Labels

We have divided the labels into logical groups according to the appropriate form of description. The basic information conveyed is its name, data-type, and some short commentary useful to the analyst.

Where the Commentary includes a † after the text (but before any list of values) have further commentary in an Additional Notes paragraph at the foot of the table.

Where Commentary is shown in "double-quotes", it equals the boilerplate text found on the most recent version of the nurse form. I have occasionally added text [between square brackets] to provide context for the question. Whilst we do not expect nurses to recite these verbatim, they may provide a guide as to how the question was put to the participant. In some cases, the question is aimed at the nurse rather than participant (e.g. asking whether certain procedures were carried out).

When the range of allowable values for a label is limited to some pre-defined set, we set out the legitimate values in **BOLD** in the Commentary.

Values that are so far from the normal range that they are considered to be "impossible" have been replaced in this extract by Unusable. The ranges considered "possible" are set out in Table 19 (page 32).

General Questions asked by the Nurse

Table 4 holds the questions entered by the nurse, plus some miscellaneous labels.

Label	Type	Commentary
barcode	NUMBER (5)	Health-screening identifier, unique within the extract. †
part_id	NUMBER (7)	Unique identifier for participant within the cohort. †
source	VARCHAR2 (3)	Database schema this barcode was extracted from: **S**: Excel-based prototype. **C**: Oracle System.
protocol	VARCHAR2 (3)	The intended protocol: **{PILOT, MAIN, FOLLOW-UP}**.
repeat_declared	NUMBER(1)	The value of this label is: † **0**: no repeat declared, no repeat found. **1**: repeat declared but not found. **2**: no repeat declared, but we found at least one. **3**: repeat declared and found.
when_screened	DATE	When the first clinic appointment took place. †

Label	Type	Commentary
when_screened_source	VARCHAR2 (9)	The source of the when_screened value: **N/A** no date available. **NURS ONLY** nurse form date only available. **ECG ONLY** ECG date only available. **ECG** ECG date used. **NURS** nurse form used
timer_start_nscr	DATE	The earliest date-time recorded by the nurse-form. This will normally be within a few minutes of the participant sitting down with the nurse.
nurse_id_first	NUMBER (38)	An arbitrary identifier for the nurse with whom the participant had their first clinic appointment. We have merged nurse-names that were misspelt.
force_id	NUMBER (38)	An arbitrary identifier for the force in which the participant is serving at the time of the screen, as reported by the participant.
results_sent_to	VARCHAR2(4)	"To whom should the participant's results be sent?" † **ME** Participant only **GP** GP only **MEGP** Participant & their GP **NONE** Neither.
is_urine_sample_given	YESANY	"Has the participant provided a urine sample?"
hours_since_last_urine	NUMBER	"When did you last pass urine previous to giving this sample?" This question, measured in hours, was dropped after the Pilot.
hours_since_eat_blood	NUMBER	How many hours passed since the volunteer eat or drank anything other than water? †
pulse_wave_done	YESANY	"Have you performed a pulse-wave measurement?"
food_diary_received	YESANY	"Please check box if [Food Diary] brought with"
has_drunk_within_24hours	YESANY	"Has the participant drunk any alcohol in the last 24 hours?"
airwave_diary_typical	VARCHAR2 (4)	Is the usage reported in the Airwave Usage Diary typical of an average week? **YES** "Yes" **MORE** "No. I usually use the radio MORE" **LESS** "No. I usually use the radio LESS" **NONE** "Diary not completed" **NOTU** "Diary not completed (not a user)" (only available since July 2011).

Label	Type	Commentary
is_menses	YESNO	"Is the participant menstruating today?" For males, we report Not Applicable.
is_pregnant	YESNO	Whether the participant is pregnant today. For males, we report Not Applicable.
outbound_postcode	VARCHAR2 (4)	The outbound portion of the postcode given by the participant as their home address.
beer_drunk	NUMBER (3)	Quantity of beer drunk (see also: beer_type) "during a typical week"
beer_type	VARCHAR2 (5)	Units of beer_drunk: **{PINTS, UNITS}**
wine_drunk	NUMBER (3)	Quantity of wine drunk (see also wine_type) "during a typical week".
wine_type	VARCHAR2 (7)	Units of wine_drunk: **{BOTTLES, UNITS}**
other_alcohol_units	NUMBER (3)	Spirits / Other units of alcohol drunk "during a typical week".
is_smoker	YESNO	"Does the participant smoke?"
cigarettes_per_day	NUMBER	"This section relates to the participant's normal consumption of cigarettes. You may ignore the very occasional cigar or pipe-tobacco." "About how many cigarettes do you smoke per day?"
smoking_age_started	NUMBER	"If you do smoke now, what age did you start?"
ecg_done	YESNO	"Was an ECG performed?"
when_ecg_first	DATE	When the first ECG was performed, as recorded by the ECG machine itself.
when_ecg_last	DATE	When the last ECG was performed.
ecg_result_summ_reported	VARCHAR2(20)	This is the confirmed ECG interpretation of the clinical significance of the ECG, in the form of a short phrase. It is returned by Glasgow CARE some weeks after the clinic. †

Table 4: Nurse Collected Data

Additional Notes for Table 4

barcode

Each barcode represents one complete screen. The value of the barcode is assigned in an arbitrary manner at the clinic, and although participants from a given police force are usually found within certain barcode ranges, the barcode contains no other information about the participant.

part_id

part_id values are assigned in sequence as they arise. They are not unique within the extract.

protocol

The protocol reported is derived from the data collected for the barcode. In general, all barcodes processed by the Excel system are **PILOT**, whereas barcodes processed on the Oracle System are **MAIN**. We can differentiate between the two based on when the screen took place, which measurements were taken (types of file collected) and the format of the feedback letter sent to the participant. Where data is present from both studies, we determined the protocol based on whether the data is mainly one or other. In any case, the extract contains all the data that was collected on the barcode, whether or not it is expected in the protocol. We carried out a follow-up study of existing participants beginning in November 2015. Barcodes resulting from the follow-up have protocol value of **FOLLOW-UP**.

repeat_declared

We aimed to offer every twentieth person a repeat screen, which would take place at least one month after his or her initial screen. In practice, somewhat fewer than this actually attend for a repeat appointment, although some participants who were not offered a repeat arrange to have a second screen anyway.

At the clinic, the nurse asks participants if this is a repeat screen and the barcode of the original screen, if remembered. No checking of this response is made at the clinic. The database independently determines whether the screen is a repeat by identifying a common part_id.

We count a repeat has having been correctly declared even if the original barcode supplied by the nurse turns out to have been wrong.

when_screened

This is derived as the earlier of the timestamp of the ECG machine when it was taken, and the timestamp on the laptop on successful submission of the nurse-form.

results_sent_to

The NONE option was removed early in the Study as virtually no participants selected this option deliberately, although nurses occasionally submitted it by mistake. It is still used internally to handle exceptional cases (e.g. death of participant before feedback returned).

hours_since_eat_blood

Where source = **C**, this field was collected by asking for a clock time (9am, 10pm etc.) and we have converted the value to a duration by comparing it with relevant nurse-form timestamp relevant to taking blood. When hours_since_eat_blood < 0 we added 24h, so that a barcode screened at 8am who reports 11pm as their last time of eating will have hours_since_eat_blood = 9. This approach is obviously poor, so we replaced it early in the Study with a direct question asking for the number of hours since last eating.

For participants who returned for a rebleed after an incomplete venepuncture, the value returned would be Unusable unless the number of hours recorded happened to be identical on each visit. If a user requires an export that includes all the available values for this label cross referenced against laboratory-assay, please contact the Study Team.

In the pilot, the record-keeping described above may not be so detailed. Users for whom these data are critical to their analyses are recommended to check whether when_screened

is always one the day prior to each of the machine dates (Table 12). If more than one day apart, it will usually imply that a rebleed was done on a new sample for which hours_since_eat_blood is not available.

ecg_result_summ_reported

Each ECG interpretation has a summary component (a short phrase) and a set of more detailed results. We include the summary interpretation, only, in this extract, and include the ECG's detailed results in ??. There are six distinct values in the extract, and several additional ones have been added since this extract was completed.

The values in this extract are in Table 5.

Interpretation	Notes
ABNORMAL	Further clinical investigation of these abnormal results is merited. We forwarded a copy of the trace to the GP and / or participant.
BORDERLINE ABNORMAL	As for ABNORMAL.
BORDERLINE NORMAL	Although atypical, the ECG does not merit in itself merit further clinical investigation.
BRADYCARDIA – NORMAL	The participant's pulse rate was lower than normal, but the ECG itself was within normal limits.
NORMAL	Within normal limits.
TACHYCARDIA NORMAL	The participant's pulse rate was higher than normal, but the ECG itself was within normal limits.

Table 5: ECG Interpretations

ECG Detailed Results

The detailed interpretation of the ECG results include eight numeric values describing the trace geometry; up to 20 Group Codes (a system of commentary proprietary to Glasgow CARE), and up to 24 Minnesota Codes (which we understand to be a more widely used classification).

We do not use the Contingency Values within the geometry section because values appear to be legitimately negative in some cases. A missing value is therefore returned as NULL and should be interpreted as Unusable. Missing values within either the Minnesota or Group Codes section simply mean that there is no such statement (Not Applicable).

A small number of participants that had an ECG and received a summary interpretation (ecg_result_summ_reported) do not have a detailed result. This occurs when the machine-stored trace was lost at the clinic and a summary interpretation was provided from the paper trace

Label	Type	Commentary
Barcode	NUMBER (5)	Health-screening identifier.
ecg_id	NUMBER (17)	Unique identifier for the recording. †

Label	Type	Commentary
when_recorded	DATE	A timestamp from the ECG machine.
Rate	NUMBER	Pulse rate in beats per second.
p_axis	NUMBER	P axis.
qrs_axis	NUMBER	QRS axis.
t_axis	NUMBER	T axis.
pr_interval	NUMBER	PR interval.
qrs_duration	NUMBER	QRS duration.
qt_interval	NUMBER	QT interval.
qtc_interval	NUMBER	QTC interval.
minnesota_group1_l	NUMBER(3)	Group 1 anterolateral Minnesota code. †
minnesota_group1_p	NUMBER(3)	Group 1 posterior Minnesota code. †
minnesota_group1_a	NUMBER(3)	Group 1 anterior Minnesota code. †
minnesota_group2_1	NUMBER(3)	First Group 2 Minnesota code. †
minnesota_group2_2	NUMBER(3)	Second Group 2 Minnesota code. †
minnesota_group3	NUMBER(3)	Group 3 Minnesota code. †
minnesota_group4_l	NUMBER(3)	Group 4 anterolateral Minnesota code. †
minnesota_group4_p	NUMBER(3)	Group 4 posterior Minnesota code. †
minnesota_group4_a	NUMBER(3)	Group 4 anterior Minnesota code. †
minnesota_group5_l	NUMBER(3)	Group 5 anterolateral Minnesota code. †
minnesota_group5_p	NUMBER(3)	Group 5 posterior Minnesota code. †
minnesota_group5_a	NUMBER(3)	Group 5 anterior Minnesota code. †
minnesota_group6	NUMBER(3)	Group 6 Minnesota code. †
minnesota_group7_1	NUMBER(3)	First Group 7 Minnesota code. †
minnesota_group7_2	NUMBER(3)	Second Group 7 Minnesota code. †
minnesota_group8_1	NUMBER(3)	First Group 8 Minnesota code. †
minnesota_group8_2	NUMBER(3)	Second Group 8 Minnesota code. †
minnesota_group8_3	NUMBER(3)	Third Group 8 Minnesota code. †
minnesota_group8_4	NUMBER(3)	Fourth Group 8 Minnesota code. †

Label	Type	Commentary
minnesota_group9_l	NUMBER(3)	Group 9 anterolateral Minnesota code. †
minnesota_group9_p	NUMBER(3)	Group 9 posterior Minnesota code. †
minnesota_group9_a	NUMBER(3)	Group 9 anterior Minnesota code. †
minnesota_group9m_1	NUMBER(3)	First Group 9m Minnesota code. †
minnesota_group9m_2	NUMBER(3)	Second Group 9m Minnesota code. †
group_codes_count	NUMBER(2)	The number of non-null Group Codes present in the following columns.
group_code_[01 .. 20]	VARCHAR2(11)	Glasgow Group Code(s) †

Table 6: ECG Detail

ecg_id

This is a unique ID generated by the ECG management system. It appears to be generated from the recording date, time and recording device.

minnesota_group[]

Each Minnesota Code consists of a three digit number >= 111 that corresponds to narrative interpretation. They are usually represented as three separate digits separated by hyphens; but they have been supplied to us without the hyphens, and we mirror that format.

group_code_[01 .. 20]

The group statements consist of up to twenty strings that each constitutes a Glasgow Group Code. Each Group Code is made up from three values separated by hyphens: group number (2 digits), statement type (2 digits) and statement code (5 digits). Taken together, they form a compound key within a table of narrative interpretations.

Body Composition Analysis

Data in were obtained from the Tanita machine in the clinic and entered by hand onto the nurse form. As of summer 2012, we are attempting to use an automatic data recording system that connects the nurse-form directly to the Tanita.

There are three named contraindications for Tanita: pregnancy, a pacemaker, and the presence of metal objects in the body. We also included an “other” option in case other reasons emerged during rollout.

The directly measured data are the impedance values; all else is derived from these values using an unknown algorithm buried within the machine. We asked the supplier for details but they declined to reveal their sources. Whatever the algorithms for these derived data might be, we understand that they are unreliable for those considered “athletes”.

Impedance values and derived values are Not Applicable when there is a contraindication.

Label	Type	Commentary
takes_intense_exercise	YESNO	"10+ hours / week of intense exercise?"
body_type	VARCHAR2 (8)	"Which Body Type?" **{STANDARD, ATHLETE}** †
fat_percentage	NUMBER	Percentage body fat (%)
total_body_water	NUMBER	Total body water (kg)
impedance_of_body	NUMBER	Impedance of the whole body (Ω)
impedance_of_left_arm	NUMBER	Impedance of the left arm (Ω)
impedance_of_left_leg	NUMBER	Impedance of the left leg (Ω)
impedance_of_right_arm	NUMBER	Impedance of the right arm (Ω)
impedance_of_right_leg	NUMBER	Impedance of the right leg (Ω)
imp_contra_metal_implant	YES	Whether the participant has a metal implant, which is a contraindication for this procedure.
imp_contra_pacemaker	YES	Whether the participant has a pacemaker, which is a contraindication for this procedure.
imp_contra_other	YES	Some other contraindication for the Tanita. Note that pregnancy - is_pregnant - is also a contraindication

Table 7: Body Composition Analysis

Additional Notes for Table 7

body_type

By default, a person who takes >10 hours / week of intense exercise will be deemed athletic. However, nurses can override this determination based on their subjective judgement of the individual sitting before them.

Physical Measurements

The labels in Table 8 are the measurements of physical characteristics. Two measurements normally take place for each barcode. Values recorded are numbers and are listed in Table 8. We report both measurements and compute a mean. If the mean cannot be computed because there are no valid values, "not found" is reported.

The stool used for the sitting-height measurements is 60.4 cm high, and this amount has already been subtracted from the measured value in the extract. The Sitting Height Ratio is derived from (sitting_height ÷ height). For the 2013 extract, the height of the stool used in the computation was 60.7 cm.

Body Mass Index (BMI) is computed for each trial as (weight ÷ height^2). Its mean is computed from the individual BMI values.

When we compute a derived value from any set of values where one of the values is missing, we return any contingency code that is shared by all the values; or, when there are two or more different contingencies, Unusable.

	Measurements [1..2]	Mean
Height (cm)	height_[1..2]	height
Girth of Hip (cm)	hip_girth_[1..2]	hip_girth
Weight (kg)	weight_[1..2]	weight
Sitting Height (cm)	sitting_height_[1..2]	sitting_height
Sitting Height Ratio	sitting_height_ratio_[1..2]	sitting_height_ratio
Girth of Waist (cm)	waist_girth_[1..2]	waist_girth
Waist / Hip Ratio	waist_hip_[1..2]	waist_hip
Body Mass Index (kg/m^2)	body_mass_index_[1..2]	body_mass_index

Table 8: Physical Measurements

Blood Pressure

Participants normally have three blood pressure readings, taken consecutively, each of which records systolic and diastolic blood pressures (mm Hg) and pulse (beats /min). Participants reporting themselves as diabetics have two sets of readings, one standing and one sitting; non-diabetics have the sitting measurement only, and the standing measurement will be Not Applicable.

We report all the measurements taken and compute a mean. If the mean cannot be computed because there are no valid values, we report the first meaningful contingency found on the underlying data. Standing measurement are reported Not Applicable when the participant is not already a diagnosed diabetic.

Description	Measurements [1..3]	Mean
pulse rate	bp_pulse_sitting_[1..3] bp_pulse_standing_[1..3]	bp_pulse_sitting bp_pulse_standing
Systolic blood pressure	bp_systolic_sitting_[1..3] bp_systolic_standing_[1..3]	bp_systolic_sitting bp_systolic_standing
Diastolic blood pressure	bp_diastolic_sitting_[1..3] bp_diastolic_standing_[1..3]	bp_diastolic_sitting bp_diastolic_standing

Table 9: Blood Pressure Measurements

Supporting data for the blood-pressure readings (sitting and standing) are in Table 10.

handedness	VARCHAR2 (5)	"Is the participant left or right handed?" **{LEFT, RIGHT}.**
bp_arm_used	VARCHAR2 (5)	"Which arm did you use for the [blood pressure] measurement?" **{LEFT, RIGHT, UNKNOWN}.**
bp_cuffsize	VARCHAR2 (7)	"Which [blood pressure] cuff was used?" **{REGULAR, LARGE, UNKNOWN}.**

Table 10: Circumstances of Blood Pressure Measurements

Pre-Existing Medical Conditions

These results all refer to "Current or Prior Medical Conditions Diagnosed by a Medical Doctor" and are available only for the main study. In the pilot, we asked a slightly different question that included diagnoses effecting members of the participant's family.

There are six groups of responses, each of which refer to a medical condition. There are five pre-defined conditions plus one "other" category that was introduced in later versions of the protocol.

For each condition, three questions are asked:

- Whether or not a diagnosis has been made (YESANY).
- Participant's "age at the time of diagnosis" (NUMBER (2)). If diagnosis = NO and the age is missing, we report Not Applicable. The comment can sometimes explain what the nurse was intending to report. For example, when diagnosis = NO but an age is provided, this may mean the participant was diagnosed with hypertension but is not currently considered hypertensive.
- Notes, which may or may not be present. The free-format text of up to 300 characters is redacted in this extract because of the potential for breaches of confidentiality. They are, of course, available to users of the private network.

The columns have names of the form diag_[condition], diag_[condition]_age and diag_[condition]_comments. For example: diag_stroke, diag_stroke_age and diag_stroke_comments.

The other conditions and stem labels are:

- Diabetes (diag_diabetes). In the current extract, diabetes is undifferentiated by type. From March 2011, we support additional columns for type One (diag_diabetes_type_1) and type two (diag_diabetes_type_2).
- Heart attack (diag_heartattack).
- High cholesterol (diag_high_cholesterol).
- Hypertension (diag_hypertension).
- Anything Else (diag_other_condition).

Cancers

These data were obtained from the Scottish (GROS) and English / Welsh (MRIS) cancer registries. When a participant has been flagged on the respective register, we are notified if that person has been or subsequently becomes a registered cancer patient.

The content of the data differs between registers. GROS returns more detail than MRIS, though much of it is of an administrative nature. This extract contains the fields data that are the same or sufficiently similar to be reported as a consolidated value.

Both registers are inclined to return duplicates, which we have suppressed. Uniqueness is based on clinical date, site and type. We also enforce uniqueness on cancer_number. The earliest record of the cancer is reported when duplicates are found.

MRIS has issued a small number of cancellations, which it describes as arising either from administrative errors or misdiagnoses (though the reason is not stated on a case-by-case basis). We have excluded all cancelled records.

Label	Type	Commentary
part_id	NUMBER (7)	Participant identifier (see Table 2).
cancer_number	VARCHAR2 (11)	An administrative construct that appears to be a unique identifier. When registry = **MRIS**, the number is composed of three concatenated values: a registry identifier (3 or 4 digits); the year of registration (2 digits); and, a patient identifier (5 digits). However, the value can take non-numeric form when participants' records have been shared between MRIS and GROS. This appears to be a data problem at the registry, which they are investigating.
site_code	VARCHAR2 (4)	ICD coding of the cancer's location.
type_code	NUMBER (4)	ICD coding of the cancer's type.
behaviour	NUMBER (1)	ICD coding of the cancer's morphology (histology). †
clinical_date	DATE	When the cancer was "diagnosed" or "treated".
date_type	NUMBER (1)	Interpretation of clinical_date: **1 –** Date of treatment. **2 –** "Anniversary Year". Only the year is available. **3 –** "Anniversary Date", which is the date of diagnosis. **4 –** Date Unusable.
year_registered	NUMBER (4)	Year of cancer registration.
registry	VARCHAR2 (4)	An identifier for the centre where the cancer was treated. These are digits when available. When not available, we report "GROS" or "MRIS" according to the source of the data.

Table 11: Cancer Labels

Additional Notes for Table 7

Behaviour

0	Benign
1	Uncertain whether benign or malignant Borderline malignancy
2	Carcinoma in situ: Intraepithelial / Non-infiltrating / Non-invasive.
3	Malignant, primary site
5	This is not a standard ICD code. According to the UK Association of Cancer Registries, it codes for "micro-invasive" cancers. They advise that it can be considered a variant of 3.

6	Malignant, metastatic site; Secondary site
9	Malignant, uncertain whether primary or metastatic site

Laboratory Results

These data relate to the results of blood analysis carried out at Northwick Park Institute of Medical Research (NPIMR).

Machines Used for Analysis

During the Pilot, we used the laboratory's existing elderly equipment: COBAS Mira (clinical chemistry), H1E (haematology) and ACL-300 (coagulation). For the Main Study we upgraded to Ilab 350 (chemistry); Advia 2120 (haematology); ACL-8000 (coagulation), and an Eliza Plate Reader for C-peptide (a new measurement).

Table 12 shows which machines were used for which analysis, and when. There are three values for each analyser. Times are DATE values that show the first and last uses of the analyser. The machine names are VARCHAR2 (20) fields. When the machine names or dates are not available, we report Not Applicable when the nurse had also reported that venepuncture had not been performed; otherwise, we report Not Found.

There is a degree of uncertainty concerning the machine used for analysis between May 2006 and April 2006 as we were progressively migrating from old machines to new. The Ilab was first used on 3rd January 2007, the ACL-8000 from 21st November 2006 and the Advia from ?? 2006.

	Machine Name	Date of First Analysis	Date of Last Analysis
Haematology	haematology_machine †	when_haematology_first	when_haematology_last
Clinical Chemistry	chemistry_machine †	when_clinchem_first	when_clinchem_last
Coagulation	coagulation_machine †	when_coagulation_first	when_coagulation_last
Eliza Plate Reader	-	when_cpeptide_first	when_cpeptide_last

Table 12: Laboratory Machines Used

Additional Notes on Table 12

haematology_machine

Because haematology should be performed on the day following venepuncture, we needed a fallback option in case the H1E was faulty and unable to be quickly repaired. We therefore used one of several eclectically named hospital machines. There was no need to do this for clinical chemistry or coagulation because these analyses are not time-critical.

The machine names are: **H1E, ADVIA 2120, ANGEL 2, ANGEL 3, NIGHTMARE, AND HOSPITAL.**

chemistry_machine

Only two machines have been used. When the laboratory machine was faulty, samples were stored at -20°C until the machine was repaired (usually a day or two at most).

The machine names are: **{COBAS MIRA, ILAB 350}.**

coagulation_machine

Only two machines have been used. When the laboratory machine was faulty, samples were stored at -20°C until the machine was repaired (usually a day or two at most).

The machine names are: **{ACL-300, ACL-8000}.**

Profiles

There are two distinct "profiles" (the set of measurements that we expect to be returned for each barcode) in the cohort, one for the Pilot and one for the Main Study. For the Main Study, measures were added to clinical chemistry because they were deemed interesting, and because the new chemistry machine could perform these measurements. We gained some extra results from haematology because these came from the new machine without extra effort, although strictly speaking they are not a part of the profile. To balance the budget, we had to drop some of the measures from the original profile. Table 13 shows which measurements are in which profile.

The boundary between the two profiles is far from distinct however. For some months, we ran both profiles in parallel to ensure the new machines were working correctly.

Clinical Chemistry, Eliza Plate Reader and Coagulation

The values in the following tables are all of type NUMBER.

Label	Commentary	Profile
c_reactive_protein	C-reactive protein (mg/L).	New
glucose	Glucose (mmol/L)	New
total_protein	Total Protein (g/L)	Old
albumin	Albumin (g/L)	Old
calcium	Calcium (mmol/L)	Old
sodium	Sodium (mmol/L)	Old
potassium	Potassium (mmol/L)	Old
alanine_aminotransferase	Alanine Aminotransferase (U/L)	Old
alkaline_phosphatase	Alkaline Phosphatase (U/L)	Old
bilirubin	Bilirubin (µmol / L)	Old
creatinine	Creatinine (µmol / L)	Both

Label	Commentary	Profile
total_cholesterol	Total Cholesterol (mmol/L)	Both
hdl	High Density Lipoprotein (mmol/L)	Both
gamma_gt	Gamma GT (U/L)	Both
apolipoprotein_a	Apolipoprotein a1 (g/L)	Both
apolipoprotein_b	Apolipoprotein b (g/L)	Both
urea	Urea (mmol/L)	Both
c_peptide	C-Peptide (pmol / L). Measured using the DTX 800 Eliza Plate Reader	New
fibrinogen	Fibrinogen (g/L). Reported by the coagulation machine.	Both
prothrombin_time	Prothrombin Time (seconds). Reported by coagulation machine.	Both

Table 13: Laboratory Result Other Than Haematology

Glycosylated Haemoglobin

Two sets of variables record the glycosylated haemoglobin (HbA1c) assay. From the beginning of the research until 27th May 2014, we used the Diabetes Control and Complications Trial (DCCT) method. DCCT was then superseded (manufacturer's support withdrawn) and replaced by the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) method. Each sample was measured using one method or the other (one exceptional case has both assays).

We understand that this change results from the National Glycohaemoglobin Standardisation Program, which was tackling the presence of too many false positives for elevated HbA1c.

The two methods both measure the proportion of HbA1c relative to total haemoglobin. However, the DCCT method is based on units of g/dl, whereas IFCC uses mmol/mol. The reference range of values differs between the two methods so that during the changeover, practitioners would not mistake the method used.

Because of the differences in method, the two series may not always be directly comparable. However, formulae and normal recommended ranges are published; and according to advice from our laboratory (May 2014), the new normal ranges are:

- 20-42 mmol/mol (formerly 4-6%) in non-diabetics
- 42-64 mmol/mol (formerly 6-8%) in controlled diabetics
- 64-up to 195 mmol/mol (formerly 8-20%) in uncontrolled diabetics.

They formulae below were attributed to IFCC; however, an article on Wikipedia (11th January 2017) uses slightly different coefficients.

DCCT (%) = 0.09148 × IFCC (mmol/mol) + 2.152

IFCC (mmol/mol) = 10.93 × DCCT (%) - 23.50

A comparison between the methods using a test sample resulted in IFCC values a little lower than the translated DCCT value, although no test of statistical significance was carried out.

Label	Commentary	Profile
haemoglobin_dcct	Haemoglobin measured according to the DCCT method (g/dL). In the 2013 extract, this was labelled haemoglobin_clinchem.	Both
hba1c_conc_dcct	Glycosylated Haemoglobin (g/dL). In the 2013 extract, this was labelled hba1c_conc	Both
hba1c_percent_dcct	Glycosylated Haemoglobin as percentage of total haemoglobin (%). In the 2013 extract, this was labelled hba1c_percent. We report the values computed by the analyser; or when not available: hba1c_percent_dcct = hba1c_conc_dcct ÷ haemoglobin_dcct ×100%	Both
haemoglobin_ifcc	Haemoglobin measured according to the IFCC method (mmol/mol).	New
hba1c_conc_ifcc	Glycosylated Haemoglobin (mmol/mol).	New
hba1c_ratio_ifcc	Glycosylated Haemoglobin according to the IFCC method (dimensionless). We report the values computed by the analyser; or when not available: hba1c_ratio_ifcc = hba1c_conc_ifcc ÷ haemoglobin_ifcc * 1000	New

Table 14: Glycosylated Haemoglobin (HbA1c) Results

Haematology

These results are obtained from the haematology machine and include both a general (Table 15), and differential white-cell analysis (Table 16). Strictly speaking, the haematology profile is identical across the cohort. However, the new haematology machine (Advia) returns certain extra results as standard.

The values in the following tables are all of type NUMBER.

red_blood_cell_count	Red blood cell count (106/µL)
white_blood_cell_count	White blood cell count (103/µL).
haemoglobin	Haemoglobin (g/dL). The value reported to participants is that measured by clinical chemistry
haematocrit	Haematocrit (%)
mean_red_cell_volume	Mean red cell volume (fl)
mean_cell_haemoglobin	Mean cell haemoglobin (pg).

mean_cell_haemoglobin_con	Mean Cell Haemoglobin Concentration (g/dL).
platelets	Platelets (103/µl)
red_cell_dist_width	Red cell distribution width (%). Advia only.
corpuscular_haem_conc	Corpuscular haematology concentration (g / dL). Advia only.
mean_platelet_volume	Mean platelet volume (fl). Advia only.
haemoglobin_dist_width	Haemoglobin concentration distribution width (g/dL). Advia only.

Table 15: Basic Haematology

Differential White Cell Counts

This is an analysis of the different types of white cell. The value measured is the percentage of white cells in each of six cell types. In aggregate, the sum should equal 100%, although in practice it can be off by several percentage points either way, which we consider an artefact of the measurement method.

The clinically significant number, however, is not the proportion of cells of each type but their absolute number. We report the counts in this extract, which are computed as the product of cell-percentage and total number of white cells. The haematology machine itself carries out the computation of counts, which we report where available. There is often a discrepancy in the last decimal place between the value reported and the product of white-cell-count and percentage. This is because the machine report rounded values externally, but makes computations based on internally held unrounded values.

When the administrators had to type data into a form, we reduced the risk of data-entry error by requiring them to fill in the percentage value only, leaving the computer to compute the count.

Label	Commentary
neutrophils_count	Neutrophils (10^3/µl)
lymphocytes_count	Lymphocytes (10^3/µl)
monocytes_count	Monocytes (10^3/µl)
eosinophils_count	Eosinophils (10^3/µl)
basophils_count	Basophils (10^3/µl)
large_unstain_cells_count	Large unstained cells (10^3/µl). Advia only.

Table 16: Differential White Cell Counts

Participants

Data in this table are obtained from the participants table. Values are determined by consolidating data from up to fifteen different sources (screening results, enrolment questionnaires, NHS etc.) into a single value.

Label	Type	Commentary
part_id	NUMBER (7)	Participant identifier (see Table 2).
sex	VARCHAR2 (6)	**MALE** or **FEMALE**. †
date_of_birth	DATE	Participant's date-of-birth. †
age_at_screen	NUMBER (2)	Participant's age at the time of the screen, rounded down to the nearest year.
date_of_death	DATE	Participant's date-of-death. †
cause_of_death	VARCHAR2 (4)	ICD10 code that is the "underlying" cause of death.
follow_up_status	VARCHAR2 (16)	State of follow-up on national registers. †
follow_up_date	DATE	The effective date of follow_up_status, when available.
last_registry_contact	DATE	The last date when any communication was received from the registry (GROS and / or MRIS) with which the participant was flagged.
postcode_last	VARCHAR2 (4)	The outbound portion of the participant's last reported home address.
postcode_country_enrolled	VARCHAR2 (16)	The country of residence (within the UK) at the time of enrolment. Values are England, Wales, Scotland or Northern Ireland.
postcode_country_last	VARCHAR2 (16)	As per postcode_country_enrolled, but based on the most recent address.
force_region_enrolled	VARCHAR2 (40)	The geographic region of force_id_enrolled.
force_region_last	VARCHAR2 (40)	The geographic region of force_id_last.
force_id_enrolled	NUMBER (38)	An arbitrary identifier for the force in which the participant is serving at the time of enrolment.
force_id_last	NUMBER (38)	An arbitrary identifier for the most recent force in which the participant is believed to be serving.
enrolled_source	VARCHAR2(7)	Source of force_id_enrolled as descriptive text.
last_force_source	VARCHAR2(7)	Source of force_id_last as descriptive text.
when_enrolled	DATE	Earliest date of enrolment within the cohort. †
cohort_entry	VARCHAR2(13)	The entry point into the cohort. QUESTIONNAIRE, SCREEN or BOTHONSAMEDAY

Table 17: Participants Details

Additional Notes on Table 17

sex

In the event of gender reassignment, we use the new gender in preference to the old.

date_of_birth

We report the modal value of the set obtained from the database once dates prior to 1920 have been discarded.

date_of_death

The value reported depends on the following logic, executed in the order shown.

- When date-of-death is known, its value is reported.
- When participant is apparently alive and being followed up, we report Not Applicable. When follow-up has been lost, we report Not Collected.
- When fact-of-death is known but not its date, we report Not Found.
- When there are multiple conflicting dates we report Values Conflict.

follow_up_status

We will be notified of cancers and deaths arising for participants that are registered with the NHS in England, Wales or Scotland. Their follow-up status expands on the current status of that person's follow-up.

Follow-up Status	Meaning for Participant
NHS E&W	Actively being followed up by the NHS in England and Wales (MRIS) – but for exports from 2015, see the **WITHHELD** status, below.
NHS SCOTLAND	Actively being followed up by the NHS in Scotland (GROS)
NORTHERN IRELAND	Has moved to Northern Ireland and is not being followed up there. though the fact of any death may be reported.
NO GP	Patient removed from the doctor's list by the Health Authority. Current whereabouts and status unknown.
GP UNKNOWN	NHS number traced, but unable to confirm a current Health Authority acceptance on our Central Index.
EMBARKED	Emigrated from the country and not being followed up, though the fact of any death whilst abroad may be reported.
ARMED FORCES	In the care of the military. They will be followed up by the NHS, though any cancer treatment carried out outside of a hospital in England, Wales or Scotland will not be reported.
FORCES DEPENDENT	
REJECTED	Unable to identify the person when the last request to flag was made. Usually this will be because the tracing details we submitted for the individual and those hold by the NHS are different. In due course we will resubmit these participants for flagging.

WAITING	These have not yet been submitted for flagging, usually because we have insufficient personal details to identify the person. No follow up currently.
REQUESTED	A flagging request is currently under way at the appropriate registry.
MISLAID	A flagging request was previously accepted, but the individual was not present in the most recent consolidated list of members. These will need to be investigated with the registry as they usually imply an error.
DEAD	Participant is reported to have died.
UNRECOGNISED	The most recent follow up information contained a status value we were unable to use.
WITHHELD	Because of legal difficulties, follow-up status and its corresponding date are usually missing from datasets received from NHS Digital since June 2017. We understand that this means that the participant is being actively followed-up (embarkations are still notified), but we do not have the date when they were last in contact with the NHS.

when_enrolled

This is the earliest of when_screened and the questionnaire's "Inception Date". The Inception Date is the first of the following values to be non-null: (a) the date of questionnaire signature, as recorded by the participant. (b) When the questionnaire was received for scanning by Group Sigma. (iii) When the questionnaire was loaded into the database. The date of signature and date of receipt at Group Sigma were not recorded for earlier versions of the questionnaire. We discarded implausible values for signature date.

Causes of Death

This table lists all the causes of death for dead participants. It is based on the death certificates that we receive a few months after the death is reported. This table reports only the ICD10 codes and not the free format text that is also received. The narrative is available to users of the private network.

Label	Type	Commentary
part_id	NUMBER (7)	Participant identifier (see Table 2).
Position	NUMBER (3)	The place in the death certificate (1, 2 etc.)
icd10_code	VARCHAR2 (4)	ICD10 code.

Table 18: Causes of Death

Exclusion of Outliers

Values that are so far from the normal range that they are considered to be "impossible" have been replaced in this extract by Unusable. This process has been carried out after the

normal process of quality assurance that was carried out whilst the data was being collected and validated. The ranges considered "possible" are set out in Table 19.

Variables	Possible Range
alanine_aminotransferase	0 – 400
albumin	20 – 200
alkaline_phosphatase	0 – 450
apolipoprotein_a	>0 – 20
apolipoprotein_b	>0 – 20
basophils_count	0 – 2
bilirubin	0 – 100
body_mass_index	13 – 70
bp_systolic_sitting	70 – 250
bp_diastolic_sitting	30 – 140
bp_pulse_sitting	30 – 200
bp_pulse_standing	30 – 200
bp_systolic_standing	70 – 250
bp_diastolic_standing	50 – 140
calcium	1.5 – 4.0
corpuscular_haem_conc	20 – 50
creatinine	18 – 400
c_peptide	90 – 4300
c_reactive_protein	0 – 40
diag_diabetes_age	<= age at screening
diag_diabetes_type_1_age	<= age at screening
diag_diabetes_type_2_age	<= age at screening
diag_heartattack_age	10 – age at screening
diag_high_cholesterol_age	<= age at screening
diag_hypertension_age	<= age at screening
diag_other_condition_age	<= age at screening
diag_stroke_age	10 – age at screening
eosinophils_count	0 – 9
fat_percentage	>0 – 70

Variables	Possible Range
fibrinogen	0 – 20
gamma_gt	0 – 800
glucose	>0 – 50
haematocrit	>0 – 100
haemoglobin	5 – 35
haemoglobin_clinchem	5 – 35
haemoglobin_dist_width	–
hba1c_conc	0 – 3
hba1c_percent	2 – 15
hdl	>0 – 7
height	130 – 220 †
hip_girth	BMI < 30: 0.42 – 0.90. BMI >=30: 0.45 – 1.30
hours_since_eat_blood	0 – 24
hours_since_last_urine	0 – 15
impedance_of_body	100 – 1000
impedance_of_left_arm	100 – 700
impedance_of_left_leg	100 – 700
impedance_of_right_arm	100 – 700
impedance_of_right_leg	100 – 700
large_unstain_cells_count	0 – 6
lymphocytes_count	white_blood_cell_count > 20: 0 – 40; else 0 – 15.
mean_cell_haemoglobin	15 – 45
mean_cell_haemoglobin_con	20 – 50
mean_platelet_volume	–
mean_red_cell_volume	50 – 150
monocytes_count	0 – 10
neutrophils_count	0 – 25
platelets	40 – 775
potassium	2.5 – 10.0
prothrombin_time	7 – 45
red_blood_cell_count	2.5 – 15

Variables	Possible Range
red_cell_dist_width	0 – 40
sitting_height	See sitting_height_ratio †
sitting_height_ratio	0.42 – 0.62
smoking_age_started	10 – 70
sodium	100 – 200
total_body_water	10 – 80
total_cholesterol	0 – 20
total_protein	40 – 120
urea	>0 – 40
waist_girth	BMI < 30: 0.3 – 0.7. BMI >=30: 0.4 – 1.1 For females only, exclude waist_girth when hip_girth excluded for being too low and hip_girth < waist_girth.
waist_hip	See waist_girth and hip_girth.
weight	See body_mass_index
white_blood_cell_count	1 – 50

Table 19: Included Ranges

† No exclusions are made when participant has dwarfism.

[1] System Level Security Policy, (version 1, November 2009), Andrew Heard.
[2] Oracle® Database SQL Language Reference 11g Release 1 (11.1). The function compares strings that are spelled differently but sound alike in English. It is described more fully in The Art of Computer Programming, Volume 3: Sorting and Searching, by Donald E. Knuth.