Introduction

This Annex provides an explanation of the data-fields available from the Airwave Study cancers export. Please read the Background section below, as it contains important context and caveats that you should understand before completing your analysis.

Title	Data Dictionary, Annex L
Subject	Metadata for Airwave Study cancers export
Version	1.0
Author	Heard, Andy H, Database Manager at Imperial College London. a.heard@imperial.ac.uk
Published	20/03/2024

Background

These data were obtained from the NHS cancer registries and are available to us under the terms of data sharing agreements (DSAs) with NHS England (NHSE) and NHSCR (Scotland). Because of restrictions imposed by our DSAs, use of the cancers' dataset is restricted to the narrow purposes of the Airwave Study, namely whether there is an association between use of TETRA and incident cancer. Sub-studies wishing to access the cancer dataset may apply to NHS for permission specific to their purpose. Please contact the Study team if this is relevant to you.

Cancer diagnoses are available retrospectively and prospectively from when they joined the cohort.

Two files form this export: diagnoses, and follow-up statuses of all cohort members.

Data Collection Protocols

NHSE provides data for participants with a health record in England or Wales. NHSCR serves Scotland. Data for participants living elsewhere at the time of their diagnosis, including Northern Ireland, are not available.

Until 2018, the two agencies co-operated to provide a broadly consistent record for people with health records in both nations. Then it was discovered that, for legal reasons, the necessary data sharing activities could not continue. Notwithstanding this, we have tried to produce a consolidated extract, though some inconsistency and duplications are likely to remain.

Methods of Notification

NHS has made data available to researchers according to an ever-changing protocol and varying levels of quality. However, the basic process is that we "flag" participants at one or both registries using their identifying information. Then, when any of these participants are diagnosed, the relevant details are notified to us.

Until October 2020, batches of notifications were received two to four times a year on an accumulating basis. Very occasionally we would additionally receive a consolidated report of all cancers in the cohort. In October 2020, NHSE's protocol changed to provide consolidated sets of data at every update point.

Changes to Registrations

Between successive updates of the post-2020 NHSE cancers' set, most of the records were of course unchanged. Occasionally, however, revisions to a registration were noticeable, and some registrations vanished. We aren't explicitly advised that a registration had changed or why, but this is our interpretation of the changes we have detected. General advice from the registrar is that the most accurate dataset is the most recent one.

In deciding what to present in a consolidated extract, our approach is one of full disclosure. Researchers can then assess for themselves how to handle any inconsistencies they encounter. So, for records with no change between updates, we export a rolled up version of the registration, showing the dates of its first and most recent notifications. Records not in the most recent update are flagged.

Obvious duplicates[1] and explicitly cancelled registrations are not exported.

Administrative Data

Most of the administrative data received from NHS are excluded from the extract, primarily because they are not very interesting but also because they may embed identifiers. However, we have retained a version of the registration number. The registrars' intention is that each diagnosis should retain the same registration number, even if changes occur to the diagnostic detail.

For confidentiality reasons, we have provided not the registration number itself but a hashed version. This makes for a long (40-character) variable, which we may simplify in later versions.

Interpretation of Site Code

The site-code variable of each cancer registration is an ICD code that was provided without lookup text. Most are ICD-10 codes that are easy to locate in standard tables, and ICD-9 codes (provided for historic diagnoses) are clearly distinguishable.

However, as anyone who has worked with historical ICD codes will know, there are sometimes "misses" in lookup tables. The ICD schema changes over time, and we do not, alas, have access to a full set of lookup tables of all changes that have occurred over the years. Moreover, the version of ICD applicable to any particular registration is not stated.

As an optional assist, we present our interpretation of each site-code. You may or may not find it useful. If you think that, as a domain expert, your interpretations are more accurate or more useful than ours, we would be pleased to discuss your ideas for their inclusion in a future extract.

ICD Lookup Text

We have provided two lookups for site-code based on tables downloaded from open sections of websites run by US CENTERS FOR DISEASE CONTRL AND PREVENTION (CDC) and World Health Organisation (WHO).

- The "2013" version is based on ICD-9 codes c.2009, and ICD-10 codes c.2013.
- The "2024" version is based on the April 2024 release.

Because a match is not always obtained by looking up the supplied code, we provide the code supplied by NHS, the code we matched to, and its textual description.

Behaviour

The cancer behaviour variable is a single digit that translates as per Table 1.

[1] Duplicates are defined as the same site-code, type-code, clinical date and registration number.

Code	Interpretation
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 and can be considered a variant of Code 3.
6	Malignant, metastatic site; Secondary site
9	Malignant, uncertain whether primary or metastatic site

Table 1: Behaviour Codes

Cancer Variables

The current file version is cancer-diagnoses-v2.tsv. It has 5,879 records and its CRC32 checksum is 01842AC2. Table 2: Cancer Fields describes each of the variables.

Label	Data Type	Description
part_id	INTEGER	Pseudonymous identifier that is unique for each participant.
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.
when_enrolled	DATE	When the participant enrolled in the study
age_at_clinical_date	NUMBER	The participant’s age at clinical_date (see below).
registration_number_hash	STRING(40)	A hashed version of the administrative code assigned to each cancer diagnosis.
site_code_cited	INTEGER	The site of the diagnosis as given to us by NHS. This is an ICD-9 or ICD-10 code.
type_code_cited	STRING	A standard code from ICD Oncology tables, as provided by NHS.
behaviour	NUMBER(1)	See Table 1.
clinical_date	DATE	The best date to use for date of incidence. It represents either the date of diagnosis or date of treatment (registries differ over time).
when_first_cited	DATE	Month and year we were first notified of the diagnosis.
when_last_cited	DATE	Most recent update on the diagnosis.
included last_revision	YESNO	When “N” (no), it means that the diagnosis was not present in the most recent data from NHS England. Always “Y” (yes) for Scottish data.
site_code_matched_2013	STRING	ICD code we matched site_code_cited to in ICD-9 or ICD-10 tables (2013 version).
site_code_description_2013	STRING	Descriptive text for site_code_matched_2013

Label	Data Type	Description
site_code_matched_2024	STRING	ICD code we matched site_code_cited to in the April 2024 table of ICD-10 codes.
site_code_description_2024	STRING	Descriptive text for site_code_matched_2024

Table 2: Cancer Fields

Follow Up Status

The current file version is fup-status-v1.tsv. It has 53,246 records and its CRC32 checksum is 56FCAFEB. Table 3: Cohort Members describes variables in the file.

Label	Data Type	Description
part_id	INTEGER	Pseudonymous identifier that is unique for each participant.
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.
when_enrolled	DATE	When the participant enrolled in the study
age_when_enrolled	NUMBER	The participant's age at when_enrolled.
registrar_status	STRING	The most recent follow-up status of this participant – see Registrar Status, below.
last_registrar_activity	INTEGER	Month and year of most recent registrar activity

Table 3: Cohort Members

Registrar Status

The registrar status fields are:

- **NHS ENGLAND**: In active follow-up by NHS England (includes Wales)
- **NHS SCOTLAND**: In active follow-up by NHS Scotland
- **LOST TO FOLLOW-UP**: Participant was previously being followed-up by one or both registries but is currently not in active follow-up. There are many possible reasons and we are not always fully advised of the cause. Some are by opt-out, but more usually because they are outside the jurisdiction of NHS. This is not always a permanent state: many return to follow-up after returning from a period abroad, for example.
- **DECEASED**: Participant is deceased.
- **UNLINKED**: We were never able to link to this participant, usually because NHS were unable to reliably link the demographic information we have with the records they hold. It's unlikely, at this stage of the Study, that anyone in this state will be followed-up in future.