# Coding Medications

## Introduction

This appendix describes the method for coding medications. It is a description of the prescription drugs, medical treatments, miscellaneous non-prescription remedies and food supplements that participants have reported that they were taking at the time of their health screen. This extract supplements the dataset present in the end-2012 extract[1] which recorded the observations in their raw, non-coded format.

The raw data was recorded during the health screen by the nurse. Asked what medical treatments they were currently taking, participants were encouraged to bring with them to the clinic a description of any prescription medicines being taken. The scope of the data collected has *de facto* expanded to include non-prescription medicines.

The current extract was created by matching the reported treatments against a reference-set that was screen-scraped from the website of the British National Formulary (BNF). Because many of the treatments reported by participants were recorded as free-format text, we have had to use "fuzzy matching" techniques to carry out the coding. Although care has been taken to accurately link each observation with its correct entry in the BNF, we anticipate that errors will remain. Researchers using this data are asked to report any inaccuracies they find so that we can improve the dataset.

### Terminology

An ***observation*** is a record of a single medication reported for a barcode (participant). Each barcode may have many medications. An example is "Aspirin taken daily".

A ***treatment*** is the coded version of the observation. So, developing the above example, the treatment would be ASPIRIN. Although we ask nurses to record each treatment as separate observations, we do get reports like "Aspirin and paracetamol", which will be matched to two treatments.

The ***reference-set*** is the set of standard treatments that we have obtained largely from the BNF website.

The ***arborescence*** is the set of treatments in its hierarchic form. The position of the treatment within the arborescence is defined by the numeric label; for example, "10.1.3.23 ASPIRIN" places aspirin within group "10.1.3 NON-STEROIDAL ANTI-INFLAMMATORY DRUGS".

A ***medication*** is the more generic concept of the drug. So, the medication "aspirin" appears as three treatments {10.1.3.23 ASPIRIN, 2.9.6.1 ASPIRIN, 4.7.1.2.1 ASPIRIN}

An ***indication*** is a description of the purpose of the drug; for example, "HEART FAILURE". Each treatment may have many indications.

## Location and Content of the Extract

This new coded dataset can be found in the "2013_frozen\2013-extra" directory. There are two files:

- `TREATMENTS_18.tsv` is the extract itself, a single datasheet containing, for each row, a single matched treatment and all its indications (where available).
- `Results_18.xlsx` is a spreadsheet containing detailed results on the matching process. Its worksheets are:
  - `Summary`: aggregate statistics on the categorisation of treatments.
  - `Matched-value-frequency`: aggregate statistics on the frequency with which each treatment was found in the dataset.
  - `Export Detail`: an alternative version of TREATMENTS_18.tsv.
  - `Arborescence`: the BNF reference-set used for matching.
  - `Synonyms`: alternative spellings and aliases for items in the arborescence.
  - `Cleaning`: list of words removed from observations and reference set

## BNF Treatments Reference Set

The BNF reference-set we have used is composed of 6,064 treatments structured into a family-tree type hierarchy. The data is largely based on http://www.bnf.org/ and was first accessed on the 2nd October 2013. It has been supplemented by additional items as necessary. The labels may not match those used by BNF. We did attempt to obtain a consistent database from the publishers but were unable to make contact with them.

An extract of the Reference Set is shown below.

| |
|---|
| 1 GASTRO-INTESTINAL SYSTEM (Heading) |
| 1.1 DYSPEPSIA AND GASTRO-OESOPHAGEAL REFLUX DISEASE (Heading) |
| 1.1.1 DYSPEPSIA (Generic) |
| 1.1.2 GASTRO-OESOPHAGEAL REFLUX DISEASE (Generic) |
| 1.1.3 ANTACIDS AND SIMETICONE (Heading) |
| 1.1.3.1 ALUMINIUM- AND MAGNESIUM-CONTAINING ANTACIDS (Heading) |
| 1.1.3.1.1 ALUMINIUM HYDROXIDE (Molecule) |
| 1.1.3.1.1.1 ALUMINIUM-ONLY PREPARATIONS (Chaff) |
| 1.1.3.1.1.1.1 ALU-CAP® (Branded) |
| 1.1.3.1.1.2 CO-MAGALDROX (Generic) |
| 1.1.3.1.1.2.1 MAALOX® (Branded) |
| 1.1.3.1.1.2.2 MUCOGEL® (Branded) |

**Table 1: Extract from Arborescence**

Each treatment has the following properties:

- A numeric label that uniquely identifies a medication and its place in the arborescence. An example of this label is "1.1.2".
- A description (e.g. "MUCOGEL®").
- A type that we have determined, whose range of values is:
  - *Heading:* These are high level categorisations of a set of treatments and / or sub-headings.
  - *Chaff*: These are also high level categorisations of a set of treatments or sub-headings; but unlike Headings, Chaff items will not be matched to an observation. For example, the "WITH PARACETAMOL" cannot be used for matching because the molecule combined with paracetamol is unspecified.
  - *Molecule* is a type of heading that describes a molecular compound from which drugs may be derived.
  - *Generic* treatments are those where the name is in common use (e.g. aspirin) rather than supplied by a single manufacturer. Many generics have descriptions that are identical to the molecule that is their hierarchic parent (e.g. 10.1.3.7.1 DICLOFENAC SODIUM). The matching algorithm will prefer the Generic over the Molecule.
  - *Branded* drugs are sold by a single supplier. We differentiate them from generics by the presence of the ® character in the description, although this is unlikely to be a definitive criterion.

Categories were automatically assigned from the structure of the BNF arborescence so may not be 100% accurate. We invite interested researchers to advise us of inaccuracies.

## Hierarchic Structure of Treatments

As explained above, the BNF is structured in a hierarchy, like a tree. The label can be used to macro-categorize treatments according to the:

- system they act upon (for example, if a researcher is interested to know whether a participant takes a treatment acting on the respiratory system, select labels under 3;
- treatment type (for example, if a researcher is interested to know whether a participant takes a beta-blocker, select labels under 2.4);
- molecule contained in the treatment (for example, if a researcher is interested to know whether a participant takes any treatment containing paracetamol, select the labels under {4.7.1.5, 4.7.2.17.6, 4.7.2.2.3, 4.7.2.4.4}.

Supplements and over-the-counter medications that did not exist in the BNF have been added to the reference-set as Group 16 – MISCELLANEOUS.

Finally, be aware that because each medication can appear in different places in the arborescence, a single observation of, say, “antihistamines” will generate four rows in the extract {3.4.1, 4.1.1.5, 4.6.8, 12.2.1.1}.

## Matching Algorithm

This section provides an outline of the algorithm used to link observations to the reference-set.

The nurse has two methods of recording treatments.

- Selecting from a pre-defined pick-list of treatments (our preferred method);
- Entering a free-format textual description. This is necessary when the drug is not in the list, and sometimes when the nurse wants to add additional commentary.

The former are relatively easy to interpret, as each member of the pre-defined list matches a treatment in the reference-set. When processing these data, we do not need to carry out anything other than simple matching.

### Interpreting Free-Format Observations

The free-format observations will often exactly match a treatment using simple string comparison. When that occurs, we use this match. There are many reasons why it may not easily match, for example the use of shorthand and popular names for drugs, misspellings, and additional descriptive prose relating to dose, route and so on. We attempt to overcome these difficulties using “fuzzy matching”, which is a method of finding the best match between observation and reference-set when no exact match exists.

### Fuzzy-Matching Algorithms

Several different fuzz-matching algorithms were used, as outlined below. Every match between an observation and a treatment / medication has four properties which are stated in the extract. These properties are listed and explained below.

- Any transformation / alternative-presentation of the reference-set;
- Any transformation applied to the observed data before matching took place;
- The algorithm used;
- A match-score, which is a measure of confidence in the match;

### Transformations on the Reference-Set

The reference-set text may have been modified before use by the matching algorithm. Any transformation used is recorded in the bnf_transform variable, whose values are:

- *Native*: the treatment was matched in its original form.
- *Cleaned*: the treatment was cleaned by removing special characters such as digits (often a dose) and terms such as “tablet”, “mg”. For a complete list of misspellings, please refer to `Cleaning` worksheet in `results_18.xlsx`.

- *Alias*: an alternative but equivalent text was used in place of the treatment. For example, ANTIBIOTIC is an alias of ANTIBACTERIAL DRUGS. For a complete list of aliases, refer to `Synonyms` in `results_18.xlsx` using synonym_type = "Alias".
- *Misspelling*: we matched to a common misspelling. For example, CANESTAN is often misspelt CANESTEN. For a complete list of misspellings, please refer to `Synonyms` worksheet in `results_18.xlsx` using synonym_type = "Misspelling".
- *Picklist*: this is not used by the fuzzy matching algorithms.
- *Unusable*: This is not really a transformation, but a set of words and phrases that we have explicitly determined to be unusable. For a complete list of the misspellings, please refer to `Synonyms` worksheet in `results_18.xlsx` using synonym_type = "Unusable".

### Transformation of Observation

The observed value may have been transformed before being presented to the matching algorithm, similar to the transformation on the reference-set, above. The only values returned are *Native* and *Cleaned*.

### Match Scores

The match-score is a number between 0 and 100, where 0 is no-match and 100 is perfect match. Each algorithm has a minimum acceptable score, and when it finds several candidate matches, it selects the match with the highest score. Match scores are comparable only within the same match-types; so, we cannot say that a match-score of 90 reported by the word-match algorithm is better or worse than a score of 85 reported by the similarity algorithm.

### Matching Algorithms

The variable match_type specifies the algorithm responsible for determining the reported match. Its range of values is:

- *Exact Match*: the observation matched perfectly with the BNF using simple string comparison. The score for exact matches is always 100.
- *Word Match*: a medication is a substring within the observation. For example, "EUMOVATE CREAM" (an observation) word-matches with "EUMOVATE" (13.4.8.1), scoring 100; CILEST word-matches CILESTE with a score of 86. The minimum acceptable score is 85.
- *Substring*: an observation is a substring within the medication. The score is the proportion of the matched treatment found in the observation; for example, OVRANETT substring-matches OVRANETTE with a score of 89. The minimum acceptable score is 40, subject to a minimum string size of 5 characters.
- *Similar*: the observation is matched to a treatment based on its edit-distance, which is a generally accepted measure of how alike two strings are. The method is based an algorithm devised by Vladimir Levenshtein, a Russian scientist working in 1965. We considered all matches that scored at least 75 by this measure. For example, "SULFASALAZINE" is 86% similar to "SULPHASALAZINE".

- *Hopeless Words*. Observations that failed all of the matching techniques were compared to a set of words that, when found, have proved to be unusable. These "hopeless" words are {"UNKNOWN", "UNSURE", "NOT KNOWN", "?"}.
- *Null Text*. During the algorithm, cleaning of the observation reduced it to an empty string. The observation was therefore written off as unusable, and the Match Type is Null Text.

## Summary of Matching Outcomes

In the 2012 extract, the relative proportions of records by Match Type are set out in the table below.

| Algorithm | Free Format (%) | Pick List (%) | Grand Total (%) |
|---|---|---|---|
| Exact Match | 57.98 | 100.00 | 80.51 |
| Hopeless Words | 0.13 | | 0.06 |
| Null Text | 0.06 | | 0.03 |
| Similar | 15.75 | | 7.31 |
| Substring | 2.01 | | 0.93 |
| Word Match | 24.08 | | 11.17 |
| **Total Observations (N)** | **19,144** | **22,129** | **41,273** |

## Overall Result of Each Match

Each observation is matched to the reference-set, which generates a set of candidate treatments. These candidates are winnowed down according to various rules to determine a final "chosen" match or matches. The match_result variable defines the outcome of this process. Its range of values is:

- *Unique*: the observation was matched with exactly one treatment that appears at exactly one place within the arborescence.
- *Multiple Arborescence*: the observation was matched with one medication; however, the medication exists as multiple treatments in the arborescence. For example ANTIHISTAMINES is present in {"EAR, NOSE, AND OROPHARYNX", "CENTRAL NERVOUS SYSTEM", "RESPIRATORY SYSTEM"}. This type of match will have several rows per medication in the export. The variable tree_matches_ct states the number of treatments matching across the arborescence.
- *Many Accepted Matches*: the observation matched several treatments. These correspond to cases where a single observation records >1 medications. For example, "COCODAMOL 30 MGS/500MGS PARACETAMOL TAKERN TDS CURRENTLY".
- *Written Off*: the observation is unusable.

## Description of the Extract Variables

This table contains variables likely to be of interest to all researchers using medications.

| Label | Type | Commentary |
|---|---|---|
| part_id | NUMBER (7) | Unique identifier for participant within the cohort. |
| barcode | NUMBER (5) | Health-screening identifier. |
| observed_value | VARCHAR2 (400) | Original text of the medication. This is either a label describing a treatment that was one of a fixed-list of predefined treatments; or, the narrative description of the medication entered by the nurse when the treatment was not one of the predefined list. See field_type. |
| bnf_text | VARCHAR2 (500) | The BNF description of the treatment which our algorithm matched. |
| bnf_cat | VARCHAR2 (40) | The hierarchic label for the treatment whose description is bnf_text. The value of bnf_cat was based on the original dataset but has been supplemented and amended, and is therefore proprietary. |
| bnf_type | VARCHAR2(8) | A classification of the matched treatment (heading, molecule etc.). |
| indication | VARCHAR2 (500) | The purpose of giving the drug (where available). |
| match_result | VARCHAR2(21) | A status field that records the overall outcome of the attempt to match the observation. |
| chosen_matches_ct | NUMBER(2) | The number of distinct bnf_text values that were matched. Zero means that the observation has been written off as unusable. Values >1 occur when an observation has matched more than treatment. For example, “Paracetamol and Aspirin” will match twice. |
| tree_matches_ct | NUMBER(2) | The number of distinct bnf_cat values that were matched; that is, the number of times that a treatment appears within the arborescence of the BNF. It is always >= chosen_matches_ct. |

### Additional Variables

The following variables give further information on the matching process and are likely to be of interest when further checks on the matching algorithm are needed.

| Label | Type | Commentary |
|---|---|---|
| med_id | VARCHAR2() | A compound field that uniquely identifies a match. It is in the form “x.y”, where x is a unique identification number (consecutive, no gaps) for the observation, and y identifies a sequence of matches made by the algorithm. E.g. “261.2” is the second match made for observation 261. We have excluded duplicates and less fit matches from the export, so “y” is not a contiguous series. |

| Label | Type | Commentary |
|---|---|---|
| med_numb | NUMBER (2) | The order in which the participant reported their treatments to the nurse; so, med_numb = 1 is the first treatment reported, med_numb = 2 is the second, and so on. This field is blank for treatments selected from the pick-list. |
| field_type | VARCHAR2() | How the data was recorded by the nurse: either "Pick List" (the treatment was selected from a predefined list); or "Free Format" (the treatment was recorded by the nurse as unstructured text). |
| observed_transform | VARCHAR2() | The transformation, if any, of the observed text was carried out in order to match. Either "Native": the observed_value was not transformed, i.e. we matched the observation in its original form; or "Cleaned": the observed_value was matched after having been cleaned. |
| bnf_transform | VARCHAR2() | What transformation, if any, of the treatment was carried out in order to match, |
| observed_matched | VARCHAR2(500) | The string derived from observed_value that was used to obtain the match. |
| bnf_matched | VARCHAR2(400) | The string derived from bnf_text that was used to match. |
| match_type | VARCHAR2(14) | The algorithm that was used to create this match. |
| match_score | NUMBER (3) | A measure of confidence in the match on a scale from 0 (no match) to 100 (exact match). |

## Things to Consider When Using the Medication Data

Researchers need to keep in mind that the matching algorithm is automated, albeit after much training. This means that its decisions should be accurate in the large majority, but not all, cases. Especially, observations attributed to a Heading or a Molecule may be linked to system(s) that are not relevant for a specific medication for at least two reasons.

First, one molecule can be used for a wide variety of purpose but only one indication is likely to be relevant for a participant. One example of this situation is the following: the molecule "ANTIHISTAMINES" can be found in:

- "CENTRAL NERVOUS SYSTEM" group as both "HYPNOTICS AND ANXIOLYTICS" (4.1) and "DRUGS USED IN NAUSEA AND VERTIGO" (4.6)
- the "RESPIRATORY SYSTEM" group as "ANTIHISTAMINES, HYPOSENSITISATION, AND ALLERGIC EMERGENCIES" (3.4); and,
- the "EAR, NOSE, AND OROPHARYNX" group as "DRUGS USED IN NASAL ALLERGY" (12.2.1).

However, the last match (nasal allergy) is the only one likely to be relevant for the majority of participants reporting using antihistamines.

Second, the texts of the arborescence can be more or less easy to match to a string depending of their lengths. One example of this situation is the following: the "OESTROGENS" will be matched with the "OESTROGENS" (8.3.1) which belongs to "SEX HORMONES AND HORMONE ANTAGONISTS IN MALIGNANT DISEASE" (8.3) / "MALIGNANT DISEASE AND IMMUNOSUPPRESSION" (8), since both strings match exactly. However, a more relevant match may have been "OESTROGENS FOR HRT" (6.4.1.1.2) which belongs to "SEX HORMONES" (6.4) / "ENDOCRINE SYSTEM" (6).

We therefore encourage researchers to check which links are relevant for their specific treatment(s) of interest when the observation is attributed to a Heading or a Molecule.

If researchers need to review all the matches (i.e. before selection) for a specific observation, they can refer to `Export Detail` worksheet in `results_18.xlsx`. This lists all the suggested matches, not just the one(s) that were eventually chosen (refer to match_status to determine which one were duplicate, disregarded, chosen or accepted multiple matches).

## Incomplete Indications Data

At the date of writing, the indications data are incomplete and are present only for Group 2: CARDIOVASCULAR SYSTEM and Group 6.1: "DRUGS USED IN DIABETES". Researchers interested in other indications are invited to provide similar data from the BNF website to further develop the medication extract. For more information on the format of the data requested, please contact Dr AC Vergnaud (a.vergnaud@imperial.ac.uk).

A H Heard – version 1.1 (30th July 2014)
AC Vergnaud - version 1 (15/07/2014)

[1] Variables: medicine_named_[01..13] and medicine_written_[1..8] in 2013_frozen\screenings_2013.sas7bdat