Categorising UK Biobank Self-Reported Medication Data using Text Matching

Philip Duncan Appleby (p.appleby@dundee.ac.uk)
University of Dundee https://orcid.org/0000-0002-4676-620X

Natalie S Buchan
GlaxoSmithKline Research and Development

Alexander SF Doney
University of Dundee

Emily R Jefferson
University of Dundee

Software

Keywords: UK Biobank Medication Data, Self-Reported Medication, Text Matching, ATC, BNF

Posted Date: December 18th, 2019

DOI: https://doi.org/10.21203/rs.2.19116/v1

Abstract

Introduction

There are multiple potential sources of information about medications taken by participants in large population and patient cohorts such as the UK Biobank (UKBB) study, including from self-reported information and from linkage to electronic health records. At the time of writing, only self-reported medication data from UKBB participants have been released, in the form of medication codes that are assigned at the time of assessment centre interviews. Prescribing data is due to be released by UKBB in the near future. UKBB self-reported medication codes and descriptions do not have any published structure, which means that grouping medications into broader categories to contribute to clinical phenotyping is not possible. The motivation for the software project described here was to develop an automated means of classifying UKBB self-reported medications, for clinical phenotyping purposes.

Methods

We describe software tools, developed to match UKBB medication descriptions with terms from drug classification systems. The WHO's Anatomical Therapeutic Chemical (ATC) Classification System and the British National Formulary (BNF) were selected as classification systems and were matched separately against UKBB medication coded descriptions. Manual matches can be added, for the cases where matching either fails or results in ambiguity. Manual matched codes would need to be added from the target classification system.

Results

Of the 3,646 medications reported as having been used by UKBB participants, 2,935 (80.5%) were matched with ATC system codes and 3,338 (91.6%) were matched with BNF codes. In general, medications remaining unmatched after manually matching those considered important according to clinician opinion, were either over the counter medicines with general descriptions or had very low participant report counts. A case study was conducted in which genetic associations between individual medications as phenotypes versus a Blood Pressure Genetic Risk Score were found less significant than those found once the medications had been automatically grouped.

Conclusion

Use of the matching software has been proven to assist in building medication-based phenotype proxies and to increase association significance over single UKBB medication codes. The matching software developed is available for general use and should also be applicable to other classification problems where descriptive text can be matched. https://github.com/PhilAppleby/ukbb-srmed/

Background

The UK Biobank (UKBB) is a long-term resource arising from a prospective study which has recruited approximately 500,000 participants, from across the United Kingdom, with the aim of investigating the

contributions of genetics and environment to the development of disease. Genotype data for approximately 490,000 participants together with a wide range of phenotype data are available as described by Sudlow et al [1].

Self-Reported Medications

Medications taken by participants can be a valuable indicator of disease status. Medication data were captured at the time of participant assessment centre visits in two steps; participants first indicated, via a touch-screen questionnaire, if they were currently taking any of certain important classes of medication. These were blood pressure lowering drugs, cholesterol lowering drugs, hormone replacement therapy and also if they were regularly taking over the counter medications including vitamins and supplements. In the interview which followed, a trained member of staff confirmed information indicated via the touch screen questionnaire, and selected details of particular drugs taken from a list of possible alternatives. Other medications currently taken at the time of the interview were captured during the interview. “Current” did not include taking of antibiotics or other short-term medications. There is no published structure in the UKBB-assigned medication codes and no means of grouping medicines into categories based on, for example, disease area or drug class. Furthermore, a mixture of trade names and generic drug names have been included in descriptions without any direct information on their equivalences, and while formulations are given in some cases (for example, Lipitor 10mg tablet), this does not indicate dosage (how many tablets and how many times per day). The UKBB medications coding table, which list codes versus medication descriptions and participant report count, is publicly available from the UKBB data showcase [3] as a separate download. Table 1 shows an excerpt from this table, in which there are a number of groupings which could be made, for example Atorvastatin and Lipitor, respectively the generic and a brand name for the same cholesterol lowering drug, have separate UKBB codes and were reported by significant numbers of participants, of whom only 326 reported taking both.

Code	Description	Participant Report Count
2038460150	Paracetamol	100036
1140868226	Aspirin	71297
1140871310	Ibuprofen	66161
1140861958	Simvastatin	62405
1141146234	Atorvastatin	18028
1140861998	Ventolin 100 micrograms inhaler	15342
1140884600	Metformin	15331
1140881856	Salbutamol	6221
1141176832	Seretide 50 evohaler	6212
1140909786	Beclometasone	4719
1141146138	Lipitor 10 mg tablet	3870
1140926606	Salbutamol 100 micrograms spacehaler	3432
1140862382	Becotide 50 inhaler	3222
1140884654	Beclomethasone	2090
1140862148	Serevent 25mcg inhaler	1373
1141171646	Pioglitazone	1182
1141157264	Salmeterol product	305
1140862380	Becloforte 250 micrograms inhaler	284
1140855380	Salbuvent 100 micrograms inhaler	114
1140874746	Diamicron 80mg tablet	71

Table 1, UK Biobank Medication Codes (non-contiguous excerpt), sorted in descending order of report-count.

Similarly, there are two items listed in the excerpt for salbutamol plus salbuvent is also a brand name for salbutamol.

Information on medication can be used to derive a range of phenotypes as stand-alone proxies for clinical phenotypes or as additional evidence in combination with other UKBB-derived phenotypes. For example, a participant might report that a medication (for example, a long acting beta adrenoceptor agonist) is being taken currently, indicating a diagnosis of asthma. In addition, there might be a record of an ICD10 code, assigned at the time of a hospital episode, indicating an asthma exacerbation.

A recent study has used the UKBB medication codes and descriptions directly as proxies for phenotypes [4] and UKBB medication codes have been used singly in fine-grained unstructured PheWAS data [5], though these would suffer from the problem of the appearance of participants taking related medication as controls for any given medication in an analysis. In this paper we show that our method for grouping related medications under ATC or BNF classifications, can eliminate false controls.

Drug Classification Systems

Drug classification systems are published by the World Health Organisation as the Anatomical Therapeutic Chemical (ATC) [6] system and by the British National Formulary (BNF) [7].

The ATC system consists of a hierarchy in which levels defined by prefixes of the full designation string for a compound are used to group drug therapies. For example:

ATC level1: R = Respiratory System

ATC level2: R03 = Drugs for Obstructive Airway Diseases

ATC level3: R03A = Adrenergenics, Inhalants

ATC level4: R03AC = Selective beta-2-adrenoreceptor agonists

ATC level5: R03AC02 = Salbutamol

ATC codes and descriptions were sourced from the ChEMBL database [8], which is “a curated database of bioactive chemicals with drug-like properties” maintained at the European Bioinformatics Institute (EBI), for testing purposes. The ChEMBL database is available as a free download in several database formats [9].

The BNF system uses a hierarchy of chapters, sections and subsections followed by codes to uniquely identify a drug. For example:

BNF chapter 3: Respiratory System

BNF section 3.1: Bronchodilators

BNF subsection 3.1.1: Adrenoceptor Agonists

BNF code 0301011R0BPADAW: SALBUTAMOL 400 CYCLOCAPS_CAP 400MCG

BNF coding is available at the Health Informatics Centre (HIC) of the University of Dundee as part of NHS-supplied data and this paper describes HIC’s use of this resource in classifying UKBB self-reported medication data.

The project aim was to design, write and test software to facilitate classification of self-reported medications in such a way that the data can effectively contribute to defining or refining clinical phenotypes, either alone or assembled from multiple sources within the UKBB data.

Methods

Whole words and phrases from UKBB medication descriptions were matched with those from the classification systems, for the ATC system, codes were assigned at ATC level3 and, for the BNF, chapter.section.subsection codes were used. Molecule synonyms extracted from the ChEMBL database were attached to descriptions, where possible, to increase the probability of finding a match for both classification systems.

Implementation

Figure 1 summarises the method implemented. There were two text matching steps. In the first, synonyms from the ChEMBL database were attached to both Classification System and UKBB codes, where possible – this is the “Attaching Synonyms” step. In the second, descriptions of medications and their synonyms were matched to UKBB medication and descriptions and their synonyms – this is the “Assigning Classification Codes” step. Excluded words were used to stop attempted matches to common and ambiguous words occurring in medication vocabulary.

Data Preparation

Classification system data were reformatted to place code and description pairs on lines for input to the matching scripts. Table 2 shows example lines from the ATC classification system and Table 3 shows lines from the BNF.

ATC Level3 Code	ATC WHO description
N02B	Metamizole sodium combinations with psycholeptics
N02B	Aminophenazone combinations with psycholeptics
N02B	Propyphenazone combinations with psycholeptics
N02B	Paracetamol
N02B	Phenacetin
R03A	Salbutamol
R03A	Terbutaline
R03A	Fenoterol
R03A	Rimiterol
R03A	Hexoprenaline
R03A	Isoetarine

Table 2, ATC Classification Codes (excerpt)

BNF Code	BNF Description
4.7.1	Co-codamol
4.7.1	Co-codamol with buclizine hydrochloride
4.7.1	Paracetamol
4.7.1	Nefopam hydrochloride
4.7.1	Co-proxamol
3.1.1	Rimiterol hydrob_inha 200mcg ref
3.1.1	Pulmadil_inha 200mcg (300 d)
3.1.1	Pulmadil-auto_inha 200mcg (300 d)
3.1.1	Pulmadil-auto_inha 200mcg ref
3.1.1	Salbutamol_inha 100mcg (200 d)

Table 3, BNF Classification Codes (excerpt)

Synonym Dictionary Creation

Molecule synonyms were extracted from the ChEMBL database and reformatted such that each synonym in a synonym-set - defined by having a common ChEMBL molregno, which is a unique identifier for a molecule, became the key to a generated synonym set in a synonym dictionary file. To achieve this, relevant columns were extracted from the ChEMBL database **molecule_synonyms** table and sorted into **molregno** order. An example for Molregno 138562 is shown in Table 4. A synonym dictionary was built such that each synonym in a synonym group (defined by a molregno) became the key to the other synonyms in the group, as illustrated in Table 5 for the same example. Molecules with the same Molregno and Molecule Synonym but from different sources were included once as the source was not relevant.

Molregno	Molecule Synonym	Source
138562	Lacersa	TRADE_NAME
138562	MK-733	RESEARCH_CODE
138562	Ranzolont	TRADE_NAME
138562	Simvador	TRADE_NAME
138562	Simvastatin	ATC
138562	Simvastatin	BAN
138562	Simvastatin	BNF
138562	Simvastatin	FDA
138562	Simvastatin	INN
138562	Simvastatin	TRADE_NAME
138562	Simvastatin	USAN
138562	Simvastatin	USP
138562	Simvastatin hydroxy acid	OTHER
138562	Synvinolin	OTHER
138562	Zocor	TRADE_NAME
138562	Zocor Heart-Pro	TRADE_NAME

Table 4, ChEMBL Molecule Synonyms (examples)

Key	Synonym Array
138562	simvador|lacersa|ranzolont|simvastatin hydroxy acid|zocor heart pro|mk 733|138562|simvastatin|zocor|synvinolin
lacersa	simvador|lacersa|ranzolont|simvastatin hydroxy acid|zocor heart pro|mk 733|138562|simvastatin|zocor|synvinolin
mk 733	simvador|lacersa|ranzolont|simvastatin hydroxy acid|zocor heart pro|mk 733|138562|simvastatin|zocor|synvinolin
ranzolont	simvador|lacersa|ranzolont|simvastatin hydroxy acid|zocor heart pro|mk 733|138562|simvastatin|zocor|synvinolin
simvador	simvador|lacersa|ranzolont|simvastatin hydroxy acid|zocor heart pro|mk 733|138562|simvastatin|zocor|synvinolin
simvastatin	simvador|lacersa|ranzolont|simvastatin hydroxy acid|zocor heart pro|mk 733|138562|simvastatin|zocor|synvinolin
simvastatin hydroxy acid	simvador|lacersa|ranzolont|simvastatin hydroxy acid|zocor heart pro|mk 733|138562|simvastatin|zocor|synvinolin
synvinolin	simvador|lacersa|ranzolont|simvastatin hydroxy acid|zocor heart pro|mk 733|138562|simvastatin|zocor|synvinolin
zocor	simvador|lacersa|ranzolont|simvastatin hydroxy acid|zocor heart pro|mk 733|138562|simvastatin|zocor|synvinolin
zocor heart pro	simvador|lacersa|ranzolont|simvastatin hydroxy acid|zocor heart pro|mk 733|138562|simvastatin|zocor|synvinolin

Table 5, Synonym Dictionary Entries

Attaching Synonyms (as in Figure 1, Part 1)

Matching of classification system (ATC or BNF) and UKBB descriptions to synonyms in the synonym dictionary proceeds according to the following rules:

- Whole ChEMBL synonyms only are used as keys to an in-memory synonym dictionary. In other words, ChEMBL synonyms are not broken down into component sub-phrases or single words when building the synonym dictionary.
- Matching of both classification system and UKBB descriptions to synonyms descriptions is on whole phrases, three-word prefixes, two-word prefixes and single words*. The decision tree for this is as below, the match process moves from specific to less specific. Within each of the following steps all synonyms for a medication are checked, in turn.

1. Match any whole phrase in the input record? (Also referred to as an all-word match).
    a. Yes -> Attach synonym, stop matching process, output ATC, BNF or UKBB record with synonyms.
    b. No -> Continue to step 2.
2. Match any prefix word trigram in the input record? (also referred to as a three-word match).
    a. Yes -> Attach synonym, stop matching process, output ATC, BNF or UKBB record with synonyms.
    b. No -> Continue to step 3.
3. Match any prefix word bigram in the input record? (also referred to as a two-word match).

a. Yes -> Attach synonym, stop matching process, output ATC, BNF or UKBB record with synonyms
b. No -> Continue to step 4.

4. Match any single word* in the input record?

a. Yes -> Attach synonym, stop matching process, output ATC, BNF or UKBB record with synonyms.
b. No -> Output ATC, BNF or UKBB record without synonyms.

***** Single words were subject to exclusion from matching based on the list of excluded words or if the word was short (3 characters or less) or if the word was a measure – for example, 500mg, 200ml, 5l)

Assigning Classification Codes to UKBB codes (as in Figure 1, Part 2)

As the flow of data reaches the stage shown in the right-hand half of *Figure 1*, classification codes (ATC or BNF) and descriptions + their synonyms are loaded as an in-memory dictionary for matching against UKBB codes and descriptions + *their* synonyms.

The matching algorithm is similar to that described in the section "Attaching Synonyms (Figure 1, Part1)", but proceeds as follows:

- A code dictionary is built where keys from the classification system descriptions and synonyms are saved versus classification system codes, there can be multiple codes per code dictionary key and an option is provided to output multiple matched records if required, otherwise the single most common code is output as a match to any given key.
- Keys to the code dictionary can be whole phrases from the input descriptions and synonyms and any size subset of the words in the descriptions (in practice all-word, 3-word, 2-word and one-word*).
- Matching UKBB descriptions and synonyms to the coding data dictionary, is on whole phrases, three-word prefixes, two-word prefixes and single words*. The decision tree for this is as below, the match process moves from specific to less specific. Within each of the following steps descriptions and synonyms are checked in turn.

1. Match any whole phrase in the input description? (Also referred to as an all-word match).

a. Yes -> Attach the classification system code(s), stop matching process, output matched ATC- or BNF-coded record.
b. No -> Continue to step 2.

2. Match any prefix word trigram in the input description? (also referred to as a three-word match).

a. Yes -> Attach the classification system code(s), stop matching process, output matched ATC- or BNF-coded record.
b. No -> Continue to step 3.

3. Match any prefix word bigram in the input description? (also referred to as a two-word match).

a. Yes -> Attach the classification system code(s), stop matching process, output matched ATC- or BNF-coded record.
b. No -> Continue to step 4.

4. Match any single word* in the input description?

a. Yes -> Attach the classification system code(s), stop matching process, output matched ATC- or BNF-coded record.
b. No -> Output an unmatched ATC or BNF record.

The data flow, shown in *Figure 1*, was designed to be iterative and included steps for feeding in manual matches and for adding excluded words to the excluded word list by examining the list of one-word matches and deciding if matches were reasonable.

Examples of the paths to successful matches for the BNF classification system are shown below in Table 6:

UKBB Code	UKBB Description	Matched Word or Phrase	Match Path	BNF Code	BNF Description
1140868226	aspirin	aspirin	1) aspirin:A	4.7.1	Non-Opiod Analgesics
1140871310	ibuprofen	ibuprofen	1) ibuprofen:A	10.1.1	Non-Steroidal Antiinflammatory Drugs
1140861958	simvastatin	simvastatin	1) simvastatin:A	2.12	Lipid Regulating Drugs
1141164828	adcal-d3 1.5g/10micrograms chewable tablet	synervit d3	1) adcal-d3 1.5g/10micrograms chewable tablet:A 2) synervit d3:A	9.6.4	Vitamin D
1140862086	salamol 100micrograms inhaler	salamol	1) salamol 100micrograms inhaler:A 2) salamol 100micrograms inhaler:3 3)salamol 100micrograms:2 4) salamol:1	3.1.1	Adrenoceptor Agonists

Table 6, Match Paths to Successful Matches, Suffixes denote match level (:A = All words, :3 = 3-word, :2 = 2-word, ;1=1-word, numerals in match path column denote match attempt number. In examples 1-3 a simple one-word description matched All of a description or synonym, in example 4) there was one unsuccessful match before a synonym was found to match, in example 5) the single word salamol matched following 3 unsuccessful attempts to match longer phrases.

Example paths to failed matches using the BNF classification system are shown in Table 7:

No.	UKBB Code	UKBB Description	Last Attempted Match	Match Path
1	1140876592	multivitamin+mineral preparations	multivitamin	1) multivitamin,multivitamin+mineral preparations:A 2) multivitamin mineral preparations:3 3) multivitamin mineral:2 4) multivitamin:1
2	1140911732	garlic product	garlic product	1) garlic product:A 2) garlic:A

Table 7, Match paths to failed matches. In example 1) there were no synonyms involved, attempted matches progressed from the whole phrase (all-word) through to an attempted 1-word match which failed. In example 2 there was an attempt to match the whole phrase followed by an all-word synonym (of length one word), both failed to match.

Unmatched data and potential false positives

UKBB coding data were marked 'unmatched' when the match path followed, starting from full phrase match for each synonym in turn, then reducing the number of words and ending with single words, had been exhausted without finding a match in the classification system synonym dictionary.

Single word matches were subject to further scrutiny as they tend to result in matches with incorrect classification codes and matches with a greater number of classification codes.

Failed matches for those UKBB medication codes considered important, based on clinician opinion and number of participant reports, were classified manually and added to the result of automatic matching.

Single words from matches considered suspect were added to the exclusion list – similar to a stop-word list referred to in natural language processing systems such as the Natural Language Toolkit (NLTK) [10], but specific to medication term matching.

Results

In the UKBB medication data coding table, which is described in the UKBB data showcase [3] as being "The list used by clinic nurses to code medical treatments" there are 6,745 descriptions. However, in the UKBB data

showcase page for the Self-Reported Medication data-field (20003) only 3,673 are listed as having a participant report count of 1 or more, at baseline. In the phenotype data set used for the UKBB project, of which this study is a part, there were 3,646 unique medication codes in use. For the 502,642 participants having records in this phenotype data, there are 1,307,924 instances of medications reported by 368,887 individual participants.

Matching text from the ATC classification system directly with descriptions in UKBB self-reported medication data produces fewer matches due to variation in trade and generic names for the medications. For example, in the case of asthma medications generic names Beclometasone or Beclomethasone and trade names such as Becloforte and Beclomist are unassigned in the matching process, without the use of ChEMBL synonyms, but are assigned to the ATC respiratory section once synonyms are used. BNF descriptions are more of a natural match with UKBB descriptions as both describe drugs available in the UK.

For example, without attached synonyms matches for the ATC system were made for 32.8% of medications and with synonyms attached to both inputs this number rose to 80.5%. For BNF coding, these figures were 85.4% and 91.6% respectively.

Three tests were run for each classification system:

- Test 0: A baseline test, with no ChEMBL synonyms attached to either the UK Biobank or classification system input.
- Test 1: ChEMBL synonyms were attached to the classification system data only.
- Test 2: ChEMBL synonyms were attached to both UK Biobank and classification system input.

Table 8 shows match counts from the three tests for the ATC and BNF classification systems and Table 9 shows the composition of matches for Test 2:

	ATC				BNF			
Test #	Matched	%	Unmatched	%	Matched	%	Unmatched	%
0	1195	32.8	2,451	67.2	3,113	85.4	533	14.6
1	2761	75.7	885	24.3	3,314	90.9	332	9.1
2	2935	80.5	711	19.5	3,338	91.6	308	8.5

Table 8, Comparison of matching capability using ChEMBL synonyms and without. Test 0: A baseline test, with no synonyms attached to either the UK Biobank or classification system input. Test 1: Synonyms were attached to the classification system data only. Test 2: Synonyms were attached to both UK Biobank and classification system input.

Types of Matches for Test 2	ATC	BNF
All-word matches	2,591	2,767
3-word matches	2	19
2-word matches	37	93
1-word matches	303	459
Total Automatic Matches	2935	3,338
Manual matches (clinician input)	93	8
Total Matches	3028	3,346
Percentage Total Matches (Includes manual matches)	83%	91.8%

Table 9, Different types of matches for Test 2 (synonyms used for both UK Biobank and classification systems)

For both classification systems, separate lists of manually classified codes were fed into the process and merged into the automatically-generated code data before phenotype generation. Examples of manually matched items for the ATC classification system are shown in Table 10.

UKBB Code	UKBB Description	UKBB Count	ATC Code	BNF Code	BNF Description
1140923346	co-codamol	12,283	N02B	4.7.1	Non-Opioid Analgesics and Compound Analgesic Preparations
1140865010	viscotears liquid eye gel	1,231	S01X	11.8.1	Tear Deficiency Eye Lubricant / Astringent
1141168326	kliovance 1mg/0.5mg tablet	1,042	G03F	6.4.1	Female Sex Hormones & Their Modulators

Table 10, Manually matched items. UKBB Count is number of reports – this was used as a factor in determining importance of completion of a manual match.

There were more automatic matches when using the BNF classification system due to a better alignment of UKBB descriptions to descriptions of medications in use in the UK. In both cases the possibility of obtaining matches was increased with the use of ChEMBL synonyms. A large proportion of the unmatched data related to over the counter and homeopathic preparations where there is more uncertainty about intended therapeutic purpose and less specific terminology used in descriptions captured in the UKBB data (for example ‘multivitamins’ or ‘garlic product’). Excluded words for self-reported medications are listed in Supplementary Materials, these are words which are disallowed from being in a one-word match.

The software written successfully assigns codes from both the ATC and BNF classification systems. Reliability of results has been verified by spot-checking results manually against both ATC and BNF internet searches, particularly for 1-word matches which were used to iteratively build the excluded words list. As a post-processing activity, following the matching data flow shown in Figure 1, matched codes were combined with UKBB medication data from reports by participants. Table 11 shows counts obtained from this stage.

Description	Count
Total number of people in the UKBB phenotype file used for this project.	502,642
Total number of people reporting taking medications	367,887 (73.19%)
Total number of medication reports	1,307,924
ATC medication reports match count	1,053,799 (80.57%)
BNF medication reports match count	1,137,316 (86.96%)

Table 11, Overall match counts for Self-Reported medications using both the ATC and BNF classification systems

The automatically matched list for ATC is included in the Supplemenary Materials.. Codes have been assigned at ATC level 3.

Use Case

Following code grouping and assignment of codes to participant reports two PheWAS experiments, one using UKBB codes and one using assigned BNF codes at chapter.section.subsection level as generated phenotypes, were run and the results plotted as illustrated in Figure 2, Figure 3, Figure 4 and Figure 5, note that the software used to build the PheWAS plots was R PheWAS [11], with the supplied PheWAS information and map tables substituted with codes and descriptions. The PheWAS phenotypes were built naively – participants reporting a medication with a UKBB code or with an assigned classification system code were cases, everyone else were controls with no exclusion for related medications or medication groups. The genetic instrument used in association testing was a systolic blood pressure genetic risk score, for 487,409 UKBB participants, built from a list of SNPS identified by Caulfield et al [12] in a large-scale blood-pressure study. Results obtained demonstrate the value of grouping related medicines as proxies for clinical phenotypes.

In Figure 2, associations with medications normally prescribed for hypertension (Bendroflumethiazide and, separately, Bendrofluazide (which is an older name for Bendroflumethazide), Atenolol, Lisinopril, Ramipril, Enalapril, Perindopril) show as highly significant, with the lowest p-value being approximately 10exp-100. Figure 3 shows a plot of the same data, but with higher “hits” (lowest p_values) removed, in order to scale up the lower associations.

In Figure 4 and Figure 5, the individual medications from Figure 2 have been rolled up into BNF-code-defined categories, with higher “hits” removed for scaling purposes in Figure 5. For the medication list mentioned in the section on figure 2, the three most significant drug classes are Renin-Angiotensin System Drugs at p-value 10 exp-205, Thiazides and Related Diuretics at p-value 10 exp-110 and Beta-Adrenoceptor Blocking Drugs at p-value 10 exp-85. These plots also have the visual advantage of grouping and colour-coding data into BNF chapters (such as “Cardio-Vascular System” or “Central Nervous System”).

There are two factors to consider when explaining why associations are stronger in the coding groups experiment illustrated in Figure 3. The grouping together of related drug codes not only strengthens the signal from the data by boosting numbers but also excludes participants from being counted as controls for closely related medications. Table 12 shows report counts (total numbers of participants) and PheWAS case and control counts for the individual medication PheWAS analysis. BNF classification codes and descriptions, used in the classified data analysis, are also shown to act as a visual link to Table 13. Note that Bendroflumaethaizide and Bendrofluazide are different names for the same medication.

Medication Name	Rpt Count	Cases	Controls	BNF	BNF Description
Bendroflumethiazide	28,444	27,659	306,296	2.2.1	Thiazides And Related Diuretics
Bendrofluazide	2,834	2,691	331,264	2.2.1	Thiazides And Related Diuretics
Lisinopril	14,122	13,725	320,230	2.2.5	Renin-Angiotensin System Drugs
Ramipril	24,097	23,416	310,539	2.2.5	Renin-Angiotensin System Drugs
Enalapril	3,824	3,711	330,244	2.2.5	Renin-Angiotensin System Drugs
Perindopril	7,234	7,055	326,900	2.2.5	Renin-Angiotensin System Drugs
Atenolol	19,424	18,782	315,173	2.4	Beta-Adrenoceptor Blocking Drugs

Table 12, PheWAS data for individual medications

Table 13 shows PheWAS association summary data resulting from analysing classified data – in this table the Meds counts is the number of medications comprising the classification. The effect of grouping data in this way is both to increase the number of cases and reduce the number of controls reported.

BNF Code	BNF Description	Meds	Count	Cases	Controls
2.2.1	Thiazides And Related Diuretics	20	33560	32388	301567
2.2.5	Renin-Angiotensin System Drugs	99	109556	87239	246716
2.4	Beta-Adrenoceptor Blocking Drugs	173	78522	59654	274301

Table 13, PheWAS data for classified medications

Table 14 shows the medications included in the “Thiazides and Related Diuretics” classification – the last 11 entries in this list did not generate summary data in the UKBB code test (Figure 2 and Figure 3) as they represent < 20 cases for each but would be rolled up into the classification in the BNF-coded case.

UKBB Code	UKBB Description	Rpt Count
1141194794	bendroflumethiazide	29,297
1140866122	bendrofluazide	2,845
1140866078	indapamide	1,801
1141194800	bendroflumethiazide+potassium 2.5mg/7.7mmol m/r tablet	117
1141146378	natrilix sr 1.5mg m/r tablet	115
1140866450	bendrofluazide+potassium 2.5mg/7.7mmol m/r tablet	61
1140910442	bzt - bendrofluazide	59
1140866162	hydrochlorothiazide	52
1140909706	chlortalidone	44
1140866108	xipamide	13
1140866092	metolazone	13
1140866146	hygroton 50mg tablet	12
1140866446	neo-naclex k m/r tablet	6
1140866136	neo-naclex 5mg tablet	6
1140866144	chlorthalidone	4
1140866072	hydroflumethiazide	3
1140866110	diurexan 20mg tablet	2
1140888922	nindaxa 2.5mg tablet	1
1140866094	metenix-5 tablet	1
1140866090	methyclothiazide	1

Table 14, Thiazides and Related Diuretics, List of UKBB coded medications

Discussion

Assessing the usability of code matching results depends on the use to which the data will be put. For example, a study might consider that only all-word matches to either original descriptions or synonyms, plus manual matches, are acceptable for use in phenotype generation. In our usage of the resulting match data all all-word, three-word and two-word matches were accepted. One-word matches were checked using internet searches and were clinician-verified where doubt remained, single words which did not produce acceptable matches were added to the excluded words list. The summary results shown in Table 8 and Table 9 were arrived at after numerous generation -> verification -> accept / reject cycles. It could be argued that a line should be drawn at a selected participant report-count and to only work with medications above the threshold. For example, of the 3,646 medications in use only about 50% (1,819) had a count greater than 10.

It is probable that the matching software written for this project can be used for other code conversion problems, in particular where synonyms are available. In other cases, an excluded-words list related to the vocabulary of the domain would have to be built.

Conclusion

The medication matching software pipeline produces results for code matching, which allow classification of UKBB self-reported medications and generation of medication-based clinical phenotype proxies. In general,

matching results occur sooner in the matching path followed by the matching code for drugs prescribed for major conditions as medication names are clearly defined for these. For over the counter medicines matches are less clear due to descriptions being more general, as explained in the previous section. Manual intervention can be made to fill in matched data gaps where a clinical expert considers unmatched medications important for a given study. It is anticipated that the main use of this software will be to enable a range of medication-based phenotypes to be generated for use in PheWAS.

Excluded words are included in supplemental data, as are lists of matched and unmatched UKBB codes and descriptions for both the ATC and BNF classification systems.

Declarations

Ethics approval and consent to participate

Not Applicable (all data presented is either test data or from the public domain)

Consent for publication

Not Applicable

Availability of data and material

Not Applicable

Competing interests

The authors declare that they have no competing interests

Funding

P Appleby PhD, Medical Research Council CASE Studentship award number 1577301

Authors' contributions

PDA planned, designed and wrote software for the UKBB SR medication project. NB, ERJ and ASFD provided suggestions and comment during the preparation of this manuscript.

Acknowledgements

The authors would like to thank UK Biobank participants and staff. Self-reported medication phenotype data and genotype data used in presented results were provided under UKBB Access Application #20405.

Thanks are also due to Alex Gutteridge, Toby Johnson and Robert Scott of GSK for their valuable suggestions throughout the project and to Yu Huang, University of Dundee for supplying the list of SNPs which were used to build the Genetic Risk Scores.

Availability and requirements

Project name: ukbb-srmed

Project home page: https://github.com/PhilAppleby/ukbb-srmed
Operating system(s): Unix and derivatives

Programming languages: Python, Bash
Other requirements: Python 2.7 or higher
License: GNU GPL

Any restrictions to use by non-academics: None

Supplementary Materials

Excluded Words List (Additional_file_1)

ATC Match List (Additional_file_2)

ATC Unmatched List (Additional_file_3)

References

1. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015;12(3):1–10.
2. Bycroft C, Freeman C, Petkova D, Band G, Delaneau O, Connell JO, et al. Genome-wide genetic data on ~ 500,000 UK Biobank participants. 2017;
3. UK Biobank medication data coding (http://biobank.ndph.ox.ac.uk/showcase/coding.cgi?id=4); Last accessed 15th July, 2019.
4. Eastwood S V., Mathur R, Atkinson M, Brophy S, Sudlow C, Flaig R, et al. Algorithms for the capture and adjudication of prevalent and incident diabetes in UK Biobank. PLoS One. 2016;11(9).
5. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. DoiOrg [Internet]. 2017;44(0):176834. Available from: https://www.biorxiv.org/content/early/2017/08/16/176834
6. British National Formulary (https://www.bnf.org/), NHSBSA Version; Last accessed 12th February, 2018.
7. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. (2017) 'The ChEMBL database in 2017.' Nucleic Acids Res., 45(D1) D945-D954.
8. EBI CHEMBL database (https://www.ebi.ac.uk/chembl/), Chembl version 23 (DOI: 10.6019/CHEMBL.database.23); Last accessed 14th June, 2018.
9. Natural Language Toolkit (NLTK) stop words (https://gist.github.com/sebleier/554280) Last accessed 19th February, 2018
10. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics [Internet]. 2014;30(16):2375–6. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/24733291

11. Caulfield MJ et al. Genetic analysis of over one million people identifies 535 new loci associated with blood pressure traits. Nat Genet. 2018;50(10):1412–25.

Figures

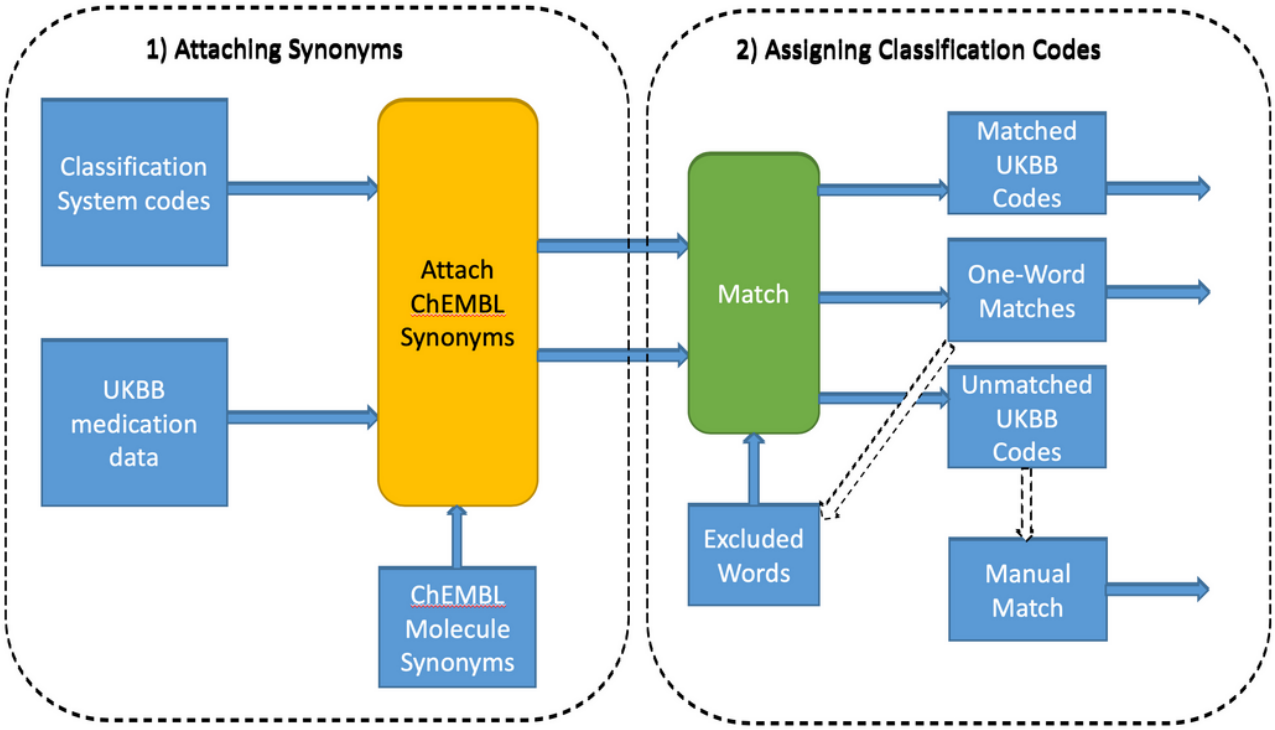

Figure 1

Method Overview. There are two main steps, both use text matching. 1) Attaching synonyms; ChEMBL molecule synonyms are attached to both Classification System and UKBB codes and descriptions. 2) Assigning Classification Codes. This is the main text matching step in which UKBB descriptions and their synonyms are compared to classification system descriptions and their synonyms. The dotted arrows show two manual feedback paths for the exclusion of common or ambiguous words and for the addition of manual matches.

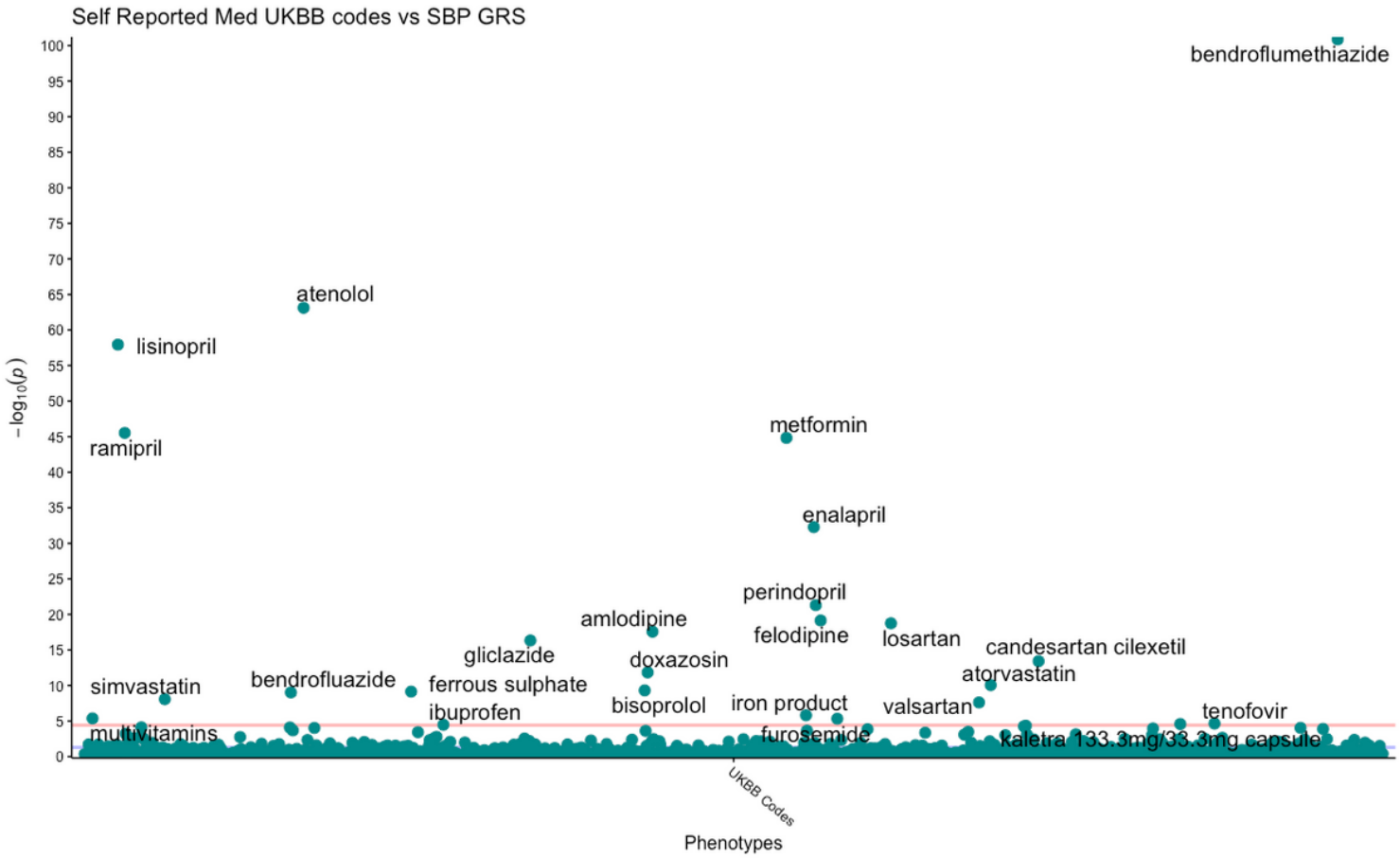

Figure 2

Plot of associations between reports of individual medications assigned to participants and a Systolic Blood Pressure Genetic Risk score (approximately 670 SNPs).

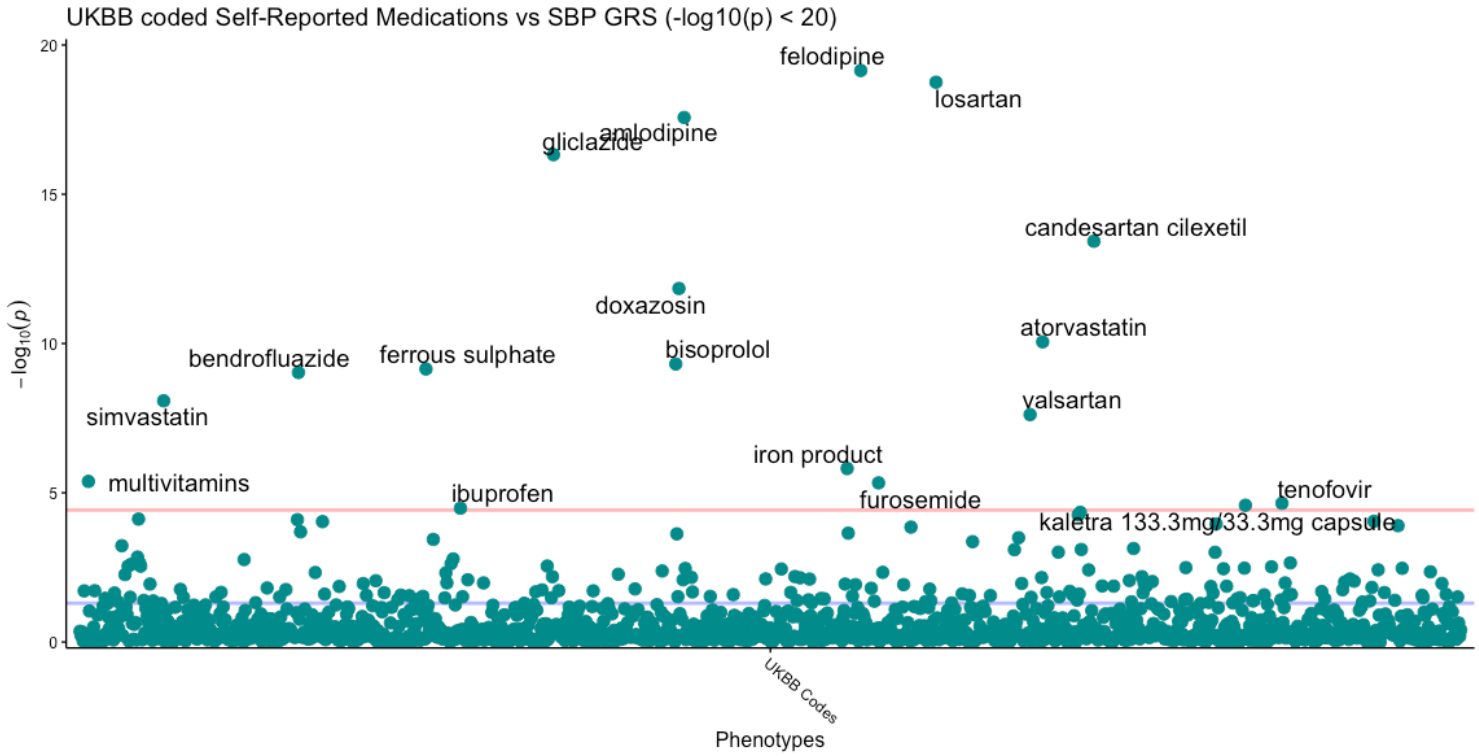

Figure 3

Plot of associations between reports of individual medications assigned to participants and a Systolic Blood Pressure Genetic Risk score (approximately 670 SNPs), with lowest p_value points removed

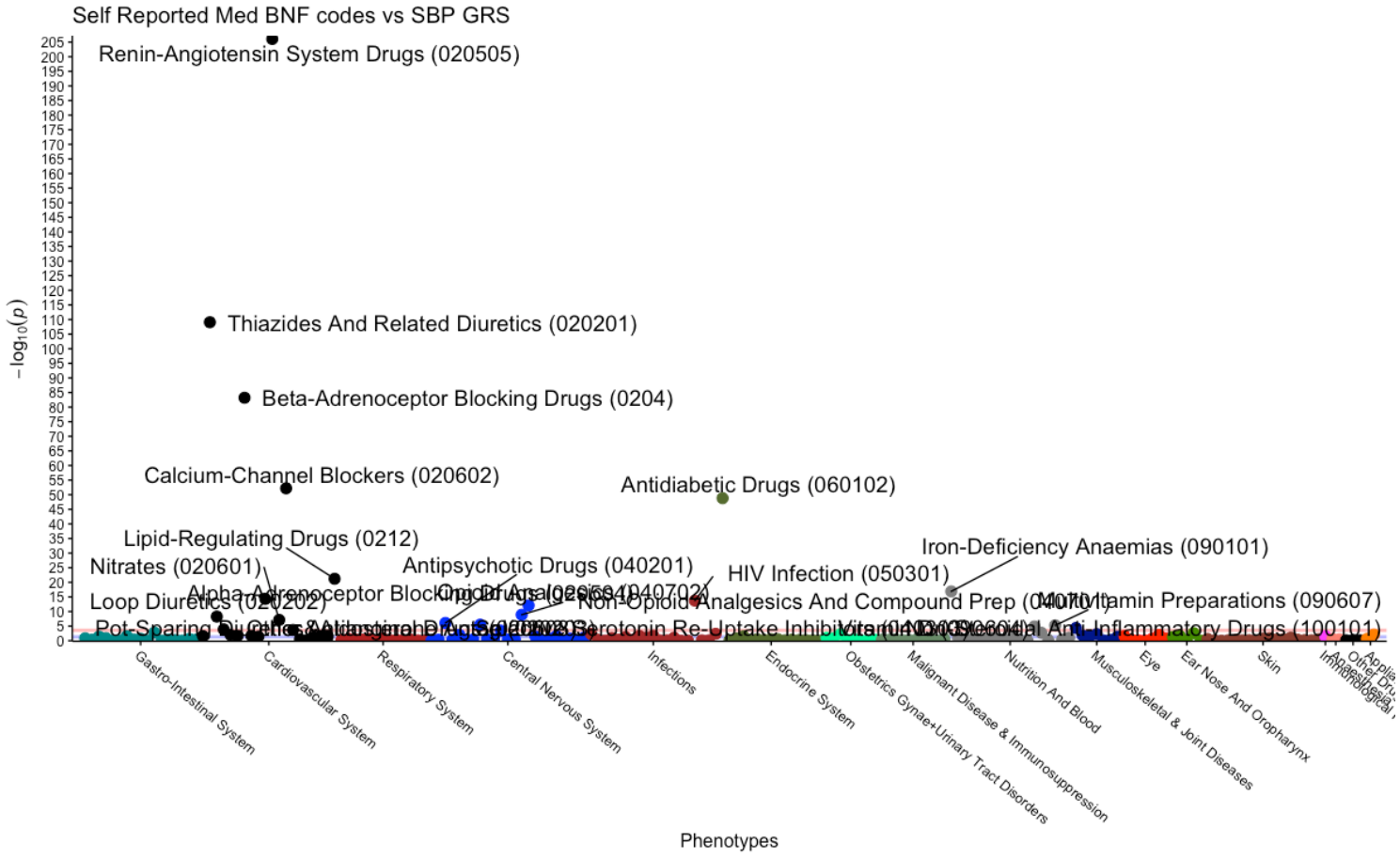

Figure 4

Plot of associations between BNF coding groups and a Systolic Blood Pressure Genetic Risk score (approximately 670 SNPs). X-axis phenotypes correspond to BNF chapters

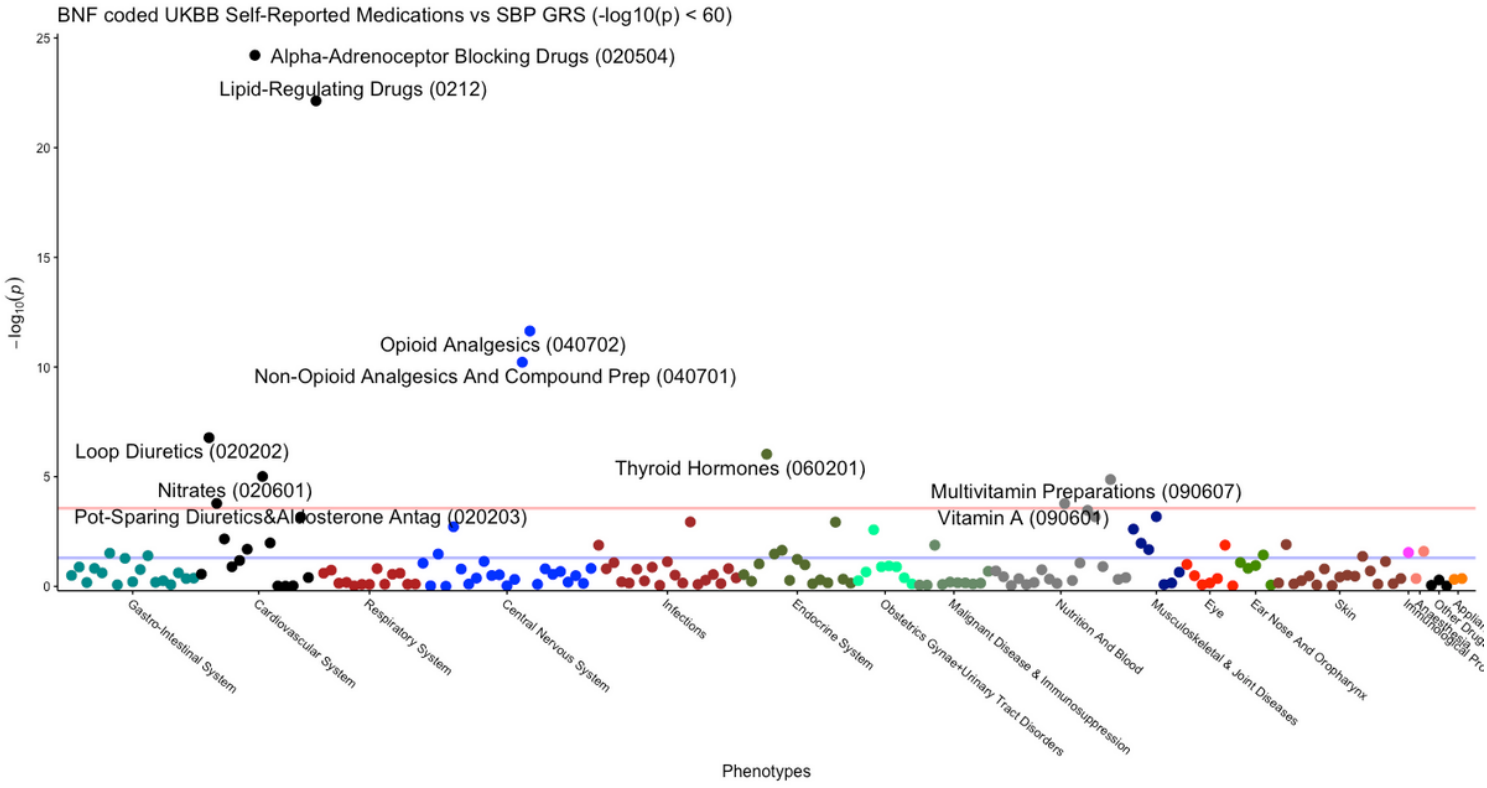

Figure 5

Plot of associations between BNF coding groups and a Systolic Blood Pressure Genetic Risk score (approximately 670 SNPs). X-axis phenotypes correspond to BNF chapters. With lowest p-value points removed

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.csv
- Additionalfile3.csv
- Addtionalfile2.csv