BMJ Open

OpenPrescribing: Normalised Data and Software Tool to Research Trends in English NHS Primary Care Prescribing 1998-2016

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019921
Article Type:	Research
Date Submitted by the Author:	04-Oct-2017
Complete List of Authors:	Curtis, Helen; University of Oxford Department of Primary Care Health Sciences Goldacre, Ben; University of Oxford, Primary Care Health Sciences
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	General practice / Family practice
Keywords:	Prescribing, National Health Service (NHS), PRIMARY CARE

OpenPrescribing: Normalised Data and Software Tool to Research Trends in English NHS Primary Care Prescribing 1998-2016

Helen J Curtis[1], Ben Goldacre[1,*]

*Corresponding author: ben.goldacre@phc.ox.ac.uk

[1]Evidence Based Medicine DataLab
Centre for Evidence Based Medicine
Department of Primary Care Health Sciences
University of Oxford
Radcliffe Observatory Quarter
Woodstock Road
Oxford OX2 6GG

Word Count: 3,685

ABSTRACT

Objectives: We aimed to compile together and normalise England's national prescribing data for 1998-2016 to facilitate research on long-term time trends, and create an open data exploration tool for wider use.

Design: We compiled data from each individual year's national statistical publications and normalised them by mapping each drug to its current classification within the national formulary where possible. We created a freely accessible, interactive web tool to allow anyone to interact with the processed data.

Setting and Participants: We downloaded all available annual prescription cost analysis datasets, which include cost and quantity for all prescription items dispensed in the community in England. Medical devices and appliances were excluded.

Primary and secondary outcome measures: We measured the extent of normalisation of data and aimed to produce a functioning accessible analysis tool.

Results: All data were imported successfully. 87.5% of drugs were matched exactly on name to the current formulary, and a further 6.5% to similar drug names. All drugs in core clinical Chapters were reconciled to their current location in the data schema, with only 1.26% of drugs not assigned a current chemical code. We created an openly accessible interactive tool to facilitate wider use of this data.

Conclusions: Publicly available data can be made accessible through interactive online tools, to help researchers and policymakers explore time trends in prescribing.

Strengths and limitations of this study

- We processed publicly-available annual data for the whole of England's community dispensing - not a sample.
- We corrected for population size, inflation, and (where possible) drugs changing name and/or classification over time.
- We produced a free, openly accessible tool for wider use, displaying trends in items, cost, price-per-item and quantity-per-item for each product for 1998-2016, which can be updated annually.
- The tool is limited to product-level data, not individual presentations, and wide-scale correction for dosage was not possible.
- Users can also download our normalised dataset in order to carry out their own analyses.

Abbreviations

BNF - British National Formulary
BSA - NHS Business Services Authority
CNS - Central nervous system
CPI - Consumer Price Index
CPRD - Clinical Practice Research Datalink
GIS - Gastrointestinal system
NHS - National Health Service
NIC - Net Ingredient Cost
ONS - Office of National Statistics
PCA - Prescription Cost Analysis
SQU - Standard Quantity Unit

INTRODUCTION

In 2016, NHS prescribing in England cost £9.20bn [1], approximately 9% of the annual NHS budget [2]. Prescribing behaviour is expected to respond within the dynamic system of evidence-based medicine, through changing patterns of disease, innovation in medical treatments, and new evidence. Monitoring long-term time trends in prescribing is therefore useful to observe changes in practice, to provide a form of feedback to ensure there are no unexpected or undesirable changes, and to facilitate tracking and forecasting of costs.

NHS Digital publish monthly and annual prescribing datasets from the NHS Business Services Authority (NHSBSA), along with static reports on prescribing trends. However this does not allow readers to interrogate topics of interest in detail, and the large datasets can be complex to manage. We provide a service at OpenPrescribing that facilitates exploration of outliers and trends for individual general practices in NHS England, which has provided over 250,000 analyses to 50,000 users over the past year. The detailed dataset that drives this service (running to over 10m rows a month) is only available from 2010 onwards. The annual Prescription Cost Analysis (PCA) data, aggregated nationally (with no data on individual practice), and by year (with no data on prescribing changes each month) is available back to 1998. This data is freely accessible, but consists of individual files for each year of prescribing, which cannot be straightforwardly combined, and therefore does not facilitate interrogation of time trends. Additionally, identifiers for individual drugs may change name, or location within the British National Formulary (BNF), over time, making simple compilation of the data impossible.

The value of PCA data is indicated by the numerous previous studies using it to assess prescribing trends [3–5] or to detect changes in response to guidelines or safety alerts [6–8]. These studies have been focused on data for a small number of drugs, manually aggregated for each bespoke analysis; furthermore, given publishing delays for academic manuscripts the data is commonly very delayed, and readers cannot easily place the findings in context of current clinical practice or expenditure.

We therefore set out to aggregate all available PCA data into a single data frame for longitudinal analysis of trends, in a service that could be easily updated; to generate an interactive online service where any user can explore and monitor time trends in prescribing using the latest available data; and to share all resources for re-use by others as open data.

METHODS

Data sources

Every available Prescription Cost Analysis (PCA) annual dataset was downloaded from NHS Digital or National Archives, covering 1998 to 2016 [9].

Data structure

The PCA data includes all items dispensed in England by pharmacy/appliance contractors, dispensing doctors, and items personally administered by doctors, whether or not they were *prescribed* in England or other parts of the UK. Items dispensed in other settings (prisons, hospitals and private prescriptions) are excluded. Prior to 2010, the data was rounded to the nearest 100 and excluded drugs with fewer than 50 items prescribed, accounting for 0.01% of total items [10]. Definitions of key terms used in the PCA data (and NHS primary care prescribing data more generally) are given in Box 1 and a full glossary of terms is available [11].

Box 1. Glossary of prescribing data terminology.

- *Items* are functionally equivalent to prescriptions; they do not take into account the quantity (number of boxes/bottles etc.) dispensed to the same person. *Items* may vary in the *quantity* prescribed.
- *Quantity* represents the amount of a drug dispensed, with units of measurement (units/tablets/grammes/millilitres etc.) dependent upon formulation, and indicated by the *standard quantity unit* (SQU).
- *Net Ingredient Cost* (NIC) represents the basic price of the medicine, i.e. the Drug Tariff price, or, if not listed, the price published by the manufacturer or supplier. NIC may be subject to further charges and/or discounts.
- The *drug name* (also known as *presentation*) includes the brand or generic (*product*) name, formulation and strength.
- For generic presentations the *product* name will match the *chemical* name (but sometimes with a different abbreviation).
- The *chemical name* is the standard registered name for the active constituent of the medicine.
- Each drug's unique 15-character BNF code is not supplied, only the first seven characters (representing the Chapter, Section, Paragraph and Sub-paragraph - see Box 2).

Every drug presentation has a unique, 15-digit structured British National Formulary (BNF) code, an example of which is given in Box 2. These hierarchical codes imply a data schema as follows: each *presentation* of a drug has a *product* name, which may be either a brand name or the generic *chemical* name; as such, each product can be mapped to a chemical. Each chemical is a member of a Paragraph in the BNF (some of which are divided into Sub-

paragraphs, which themselves often approximate to a class of drugs). Each Paragraph belongs to a Section, which is in turn a member of a Chapter (often approximating to a system of the body, such as “Cardiovascular”).

Box 2. BNF Code Structure.

Example presentation: Tradorec XL Tablets 300mg

Chapter	Section	Paragraph	Sub-paragraph	Chemical	Product	Presentation	Generic Equivalent
04	07	02	0	40	BI	AC	AM
Central Nervous System	Analgesics	Opioid Analgesics	Opioid Analgesics	Tramadol Hydrochloride	Tradorec	Tradorec XL_Tab 300mg	*

*Generic equivalent allows matching with the strength and formulation (presentation) of the generic product (which will always have product code ‘AA’).

In the PCA data, only the first seven characters of the BNF code for each drug are supplied. Each individual presentation is then only fully described in text (such as “Lipitor_Cap 20mg”), rather than the full BNF code, which can then be imputed up to the level of *product*. Although each drug’s chemical name is also supplied, chemicals are not all unique (e.g. “Other Preparations”); names may change their spelling over time; and chemicals may move between Paragraphs, Sections and Chapters. Indeed, classifications at any level of the hierarchy can be subject to renaming, spelling change, subdivision, reorganisation and removal.

Data management, aggregation, and cleaning

All data was grouped by drug name, combining those differing only by standard quantity unit (SQU). Ultimately, following cleaning, data was grouped to product level. Medical devices/appliances and any other items in pseudo-chapter numbers above 15 were excluded.

A key user-need was to explore prescribing trends for individual members of a class of drugs over time. This required that all data was normalised, with each individual drug consistently appearing in the correct location in the data schema; i.e. all individual presentations of a chemical all mapped under that chemical; and all chemicals mapped under the correct Sub-paragraph/Paragraph (often similar to drug class) of the BNF. To achieve this consistency, we aimed to map each drug to its current position in the latest BNF dictionary, up to the level of its 11-character “product” code, through an incremental process. This is summarised below and in Figure 1.

Lacking the full BNF code, we attempted to match each drug name to a current BNF presentation. Those without an exact match (e.g. formulation variants no longer available) could sometimes be matched to a similar BNF presentation name, e.g. by finding a similar formulation or using the "fuzzy" lookup add-on for Excel and validated manually [12]. Other drug names could only be matched up to current BNF codes by using their product or chemical names. Matching at each stage was improved by disregarding capitalisation, or spacing and spelling changes (e.g. Sulphur/Sulfur); these include changes identified within the data and those occurring when many old British spellings (the "British Approved Name") were replaced with international standard names (the "Recommended International Non-Proprietary Name") [13]. Remaining drug names in the most-prescribed Chapters (1-6 and 10) were matched to current drug names manually (for example, resolving non-matches due to rearrangement of word order); any others kept original chemical name, and a proxy product name was derived from the drug name field. Full methodology for this matching process is available in our technical documentation online [14] and in Supplementary Material.

We measured the extent of normalisation of drug names and classifications, and present summary statistics on these.

Normalisation for inflation and population

Prescribing costs were corrected for inflation using the UK's annual consumer price index (CPI) figures, normalised to 2016 [15]. Number of items prescribed and costs were divided by the population each year to calculate values per thousand population, based upon mid-year population estimates for England only [16].

Interactive Analysis Tool

Having generated a normalised dataset, and a method for updating it, we then set out to implement a free, interactive online data analysis tool where any user can visually explore time trends in prescribing. This was built using Tableau Public, a freely accessible interactive data presentation platform which permits rapid prototyping; however other front-ends onto the same underlying datasets could also be implemented using open source tools such as Shiny [17] or in Python libraries such as d3 [18] with more software engineer resource. Our user-needs for the tool were as follows: the ability to display trends in items and cost, normalised for total population change and inflation; and also to calculate the average cost per item and quantity per item for each product.

Having delivered the tool, we used it to generate trends data and graphs for a range of clinical areas where prescribing trends have been previously studied and published, to demonstrate the ability of our tool to replicate and extend these works.

Data and code

The full compiled and processed PCA datasets are available online via Figshare [19], SQL code is in Supplementary File, and the Trends tool (Tableau workbook) is available to use via https://OpenPrescribing.net/pca.

RESULTS

Data compilation and overall prescribing trends

All data was successfully imported. There were 169,100 lines of data in the compiled 1998-2016 dataset (Chapters 1-15) and 169,038 in the processed data, the reduction caused by aggregation over SQU. Total items (14.8 billion), cost (£136.9 billion) and distinct drug names (22,496) remained consistent before and after data processing (Table 1, S1). The inclusion of low volume prescribing in the published datasets from 2010 caused a substantial rise in the number of distinct drugs per year, but not items or cost (Table 1). As can be seen from Table 1, the inflation-corrected cost in 2016 UK sterling equivalent for all prescribing in NHS England primary care rose from £6.3bn in 1998 to £10.1bn in 2004, but then decreased to £8.3bn in 2016. Items per 1,000 population (correcting for crude population growth) has grown from 10,180 in 1998 to 19,196 in 2016.

Table 1. Summary of processed PCA data by year (drugs in Chapters 1-15 only).

Year	Distinct count of Drug Name	Items	Items per 1000	Cost	Inflation-Corrected Cost	Inflation-Corrected Cost per 1000
1998	6,338	497.0M	10,180	£4,440M	£6,280M	£128,626
1999	6,587	513.4M	10,471	£5,011M	£6,999M	£142,746
2000	6,613	535.1M	10,868	£5,284M	£7,318M	£148,648
2001	6,754	569.2M	11,510	£5,784M	£7,914M	£160,049
2002	6,834	598.6M	12,050	£6,487M	£8,768M	£176,491
2003	6,893	630.3M	12,625	£7,113M	£9,488M	£190,035
2004	6,912	666.0M	13,268	£7,645M	£10,063M	£200,482
2005	6,907	698.8M	13,808	£7,452M	£9,609M	£189,875
2006	6,810	728.4M	14,292	£7,660M	£9,655M	£189,436
2007	7,056	771.8M	15,022	£7,810M	£9,614M	£187,112
2008	7,202	816.7M	15,762	£7,716M	£9,174M	£177,047
2009	7,401	859.2M	16,461	£7,892M	£9,176M	£175,805
2010	11,703	898.4M	17,065	£8,162M	£9,193M	£174,636
2011	11,751	931.6M	17,541	£8,101M	£8,734M	£164,457
2012	12,207	968.9M	18,112	£7,802M	£8,176M	£152,836
2013	12,318	996.2M	18,494	£7,846M	£8,022M	£148,920
2014	12,576	1,027.0M	18,908	£8,022M	£8,078M	£148,718
2015	12,875	1,043.5M	19,046	£8,403M	£8,461M	£154,444
2016	13,285	1,060.9M	19,196	£8,284M	£8,284M	£149,892
Total	***22,496***	**14,810.9M**	**284,680**	**£136,914M**	**£163,006M**	**£3,160,254**

Data Normalisation

Data was normalised using the methods described above. Of the distinct drug names in the data, 87.5% were matched exactly to a current BNF name, and a further 6.5% matched approximately (Table 2). Name changes are particularly prevalent in Chapter 3 (Respiratory), due mainly to the addition in 2004 of a space when a number of doses is

given, as is common for inhalers, e.g. “Salbutamol_Inha 100mcg (200D)” became “Salbutamol_Inha 100mcg (200 D)”.

Table 2. Number and percentage of drug names subject to changes within 1998-2016 PCA data when compared to the current BNF, by Chapter. These include changes in word order, spacing, capitalisation, abbreviation, punctuation (e.g. “Califig_(California Syr Of Figs)”/“Califig_California Syr Of Figs”), spelling (e.g. “Sulphate”/“Sulfate”), brand name (e.g. “Laxoberal_Liq”/ “Dulcolax Pico_Liq”) and formulation (e.g. “Castor Oil_”/“Castor Oil_Liq”). The total count of drug names is reduced compared to Table 1 because the same drugs can appear over multiple years but only rarely in multiple Chapters.

Current Chapter code	Name/spelling change		No change		No match		Grand Total
1		0.0%	1,041	99.1%	9	0.9%	**1,050**
2	177	7.2%	2,278	92.5%	8	0.3%	**2,463**
3	238	20.8%	893	78.0%	14	1.2%	**1,145**
4	341	8.8%	3,535	90.7%	21	0.5%	**3,897**
5	119	9.1%	1,184	90.7%	3	0.2%	**1,306**
6	203	13.2%	1,320	86.0%	11	0.7%	**1,534**
7	4	0.7%	508	90.7%	48	8.6%	**560**
8	2	0.3%	538	91.8%	46	7.8%	**586**
9	210	4.0%	4,281	81.3%	776	14.7%	**5,267**
10	89	8.5%	956	90.8%	8	0.8%	**1,053**
11	18	3.0%	532	87.2%	60	9.8%	**610**
12	16	3.4%	373	78.5%	86	18.1%	**475**
13	22	1.1%	1,772	89.5%	186	9.4%	**1,980**
14	7	3.0%	198	85.0%	28	12.0%	**233**
15	15	5.2%	229	78.7%	47	16.2%	**291**
Grand Total	1,461	6.5%	19,638	87.5%	1,351	6.0%	**22,450**

Of the distinct drug names (23,275, taking into account some drugs having multiple BNF classifications), over 91% could be matched to a current product in the BNF, with no change in code (Table 3). Less than 5% could not be matched to a current product and/or chemical code, under 1% of items prescribed. These drugs were assigned proxy product names (derived from their drug name) so that all data could be presented visually, and for those not matched to a current chemical, the original chemical name was used (mostly “Other Preparations”). However, normalisation was focused on seven of the most prescribed Chapters with the greatest medical interest (1-6 and 10). The normalisation of drugs in other Chapters could therefore potentially be improved. In particular, Chapters 9 (Nutrition) and 13 (Skin) have substantial levels of prescribing, but are complex, containing many different drug names and non-drug products such as topical applications and dietary supplements. Other

groups with a particular interest in nutrition or dermatology may wish to expand our work on manual matching: we would be happy to incorporate such amendments into our dataset. Many of the code changes and non-matches have diminished over time, as expected (Table S2).

Code changes and normalisation outputs are described in Table 3. Headers indicate the highest level in the BNF hierarchy at which drugs have been subject to code changes, e.g. "Section" indicates drug names which have not changed Chapter but have moved Section. "No product match" indicates drug names matched to a chemical (9-character BNF) but with no current matching product (11-character). "No chemical match" indicates drug names matched neither to a chemical nor product. The total count of drug names increases when separated by Chapter because four drug names currently exist in two different Chapters.

Table 3. Summary of drug code changes within the 1998-2016 prescribing datasets, also separated by (current) Chapter.

	BNF code change							
	Chapter	**Section**	**Paragraph**	**Sub-paragraph**	**No change**	**No product match**	**No chemical match**	**Grand Total**
Distinct count of Drug Name	94	52	560	203	21,258	815	293	**23,275**
% of Drugs	0.40%	0.22%	2.41%	0.87%	91.33%	3.50%	1.26%	**100%**
% of Items	0.04%	0.01%	0.84%	0.51%	97.67%	0.84%	0.10%	**100%**
Distinct count of Drug Name by Current Chapter								
Current Chapter number	**Chapter**	**Section**	**Paragraph**	**Sub-paragraph**	**No change**	**No product match**	**No chemical match**	**Grand Total**
1		4	109		1,146	6		**1,265**
2	5	3	4	19	2,446	7		**2,484**
3	1	6	19		1,132	10		**1,168**
4	42	13	173		3,831	9		**4,068**
5	1		161	116	1,241	2		**1,521**
6	4	1	2	2	1,521	10		**1,540**
7	5		12		528	28		**573**
8	1		4		579	6	1	**591**
9	1	7	6	54	4,510	507	245	**5,330**
10	27				1,041	9		**1,077**
11	1	5	5		571	28	6	**616**
12					420	44	11	**475**
13	5	13	28	12	1,847	95	30	**2,030**
14			37		195	17		**249**
15	1				254	37		**292**
Grand Total	**94**	**52**	**560**	**203**	**21,262**	**815**	**293**	**23,279**

Interactive Data Analysis Tool

We created a tool which allows anyone to explore the prescribing data, available directly at https://openprescribing.net/pca. Users can search by chemical, Paragraph, Section or Chapter to view time trends in items and costs on stacked charts, where both the overall trends and the relative contribution from each product/chemical can be seen. The cost per item and quantity per item for each product are also shown, which can assist in interpretation of trends in some cases. However, these calculations carry a "use with caution" note, as items may represent different pack sizes, and quantities cannot be reliably summed across preparations because of different strengths and formulations. The page features an accompanying video walk-through demonstrating the tool.

The tool can be used to facilitate novel research into time trends, and factors associated with changes in practice such as publication of guidelines or evidence landmarks, or changes in price. It can replicate and extend the main findings of previous papers which researched trends for different clinical areas using PCA data. For example, the antipsychotic drug switches which occurred in England following a licence restriction [20] can be replicated in the tool and the trends extended to the latest data (Figure 2a). This also shows the dramatic reductions in cost that followed the expiry of patents for risperidone and olanzapine. We also replicate antidepressant prescribing trends, previously reported up to 2010 [5], and show that how the overall use of these antidepressants has continued to rise, in particular sertraline (Figure 2b). We also replicate findings on the rise of thyroid hormones [3] and testosterone [4], where we show that prescribing of these drugs continued to rise, with a disproportionate increase in cost (Figure S1a-b). We are using this dataset and tool in our academic papers on trends and variation in NHS prescribing; we encourage others to use our dataset and tool in their own work.

The tool can also be used to complement studies performed in more detailed prescribing data such as the Clinical Practice Research Datalink (CPRD), by giving the full national picture, and giving more longitudinal data that updates with new data releases. For example, several previous publications have reported on patterns of prescribing of smoking cessation medication in The Health Improvement Network (THIN) database [21–23]. This included reporting of a possible decline in prescribing despite increased incentives for GPs introduced in 2012. We can confirm this decline and show that it continued beyond 2013 (Figure 2c). We also show that the slow decline in quinine usage following safety alerts in 2010 [24] has continued at a similar pace (Figure S1c). CPRD data contains individual patient records and

can therefore be used to assess detailed questions about treatments in specific cohorts of patients. However many labour-intensive CPRD analyses have been conducted to interrogate simple broad prescribing trends which could more straightforwardly be conducted using aggregated and normalised national data, with greater coverage of years and total population. Furthermore, for analyses interrogating national trends and responses to guidelines, in many cases a prescribing change which can only be detected in individual patients' records, and cannot be detected in national data, may not be relevant in terms of population health or the health service.

Additional tabs in the tool allow discovery of higher-level trends, including Chapter and Section trends, Sections ranked by items/cost for any selected year, calculation of the change in items/cost for each Section between any selected year to the latest year, and the top 20 Paragraphs by items and cost. The Chapter trends page, for example, shows that much of the decline in prescribing costs since the peak in 2004 (Table 1) is attributable to a drop in the cost of cardiovascular drugs (Figure 3a), and the Section trends page further shows that lipid-regulating drugs (Section 2.12) and Drugs for Hypertension and Heart Failure (Section 2.5) experienced the largest cost reductions at that time (Figure 3b).

DISCUSSION

Summary

It was possible to aggregate all PCA data from 1998-2016 and normalise for most changes in drug names and classifications. Only 87.5% of drug names matched exactly to a current BNF name and 8.7% had undergone some change in classification; however all drugs in core clinical Chapters were reconciled to their current location in the data schema. We generated an interactive online service where any user can explore time trends in prescribing broken down by product, chemical, Paragraph, Section and Chapter; this openly accessible interactive data analysis tool provides overviews and insights comparable to previous labour-intensive bespoke data analysis research projects.

Strengths and weaknesses of this study

Our tool covers the data for the whole of England's community dispensing, not a sample. We are surprised to note that this is the first project aiming to aggregate long-term trends across the entire prescribing dataset, and provide an openly accessible tool for wider use. Many drugs changed name and/or classification over time, but valid chemicals were successfully assigned to all items in Chapters 1-6 and 10, and product names were derived for every

drug, allowing maximum consistency in trends analysis. The tool is limited to product-level data due to the wide number of different presentations available.

We used items to measure prescribing volume. Quantity is generally more complex for making comparisons as there is wide variation caused by the units, which may be the number of pills or millilitres, units (such as inhalers containing multiple doses) or other unit measure. There is no comprehensive information on daily dose sizes for all drugs to allow thorough normalisation (e.g. Defined Daily Dose, DDD) and this would be even more difficult for discontinued drugs. However, users can download our BNF-normalised dataset in order to apply these calculations to a subset of drugs for a more accurate analysis of trends. Using items also has limitations, as it does not take into account number of packs prescribed per prescription, pack size or dosage. We are launching this tool publicly and will monitor user volume and user-feedback: if appropriate we will improve the tool by replicating and expanding it using bespoke software as per our other data analysis tools on OpenPrescribing.net for exploring variation in prescribing at CCG and individual practice level.

Findings in context of other research

Long-term trends in prescribing have previously been reported on a wide variety of clinical areas, using PCA data as well as other sources. These are static, not updated, and rapidly out of date. Although using CPRD allows a more detailed analysis and investigation of patient factors associated with prescribing, it takes a great deal of preparation and time to complete. Our tool can replicate some trends found in CPRD, and so may provide a useful tool for preliminary investigation of trends. It can also help to confirm whether findings from regional datasets of rich individual patient data (IPD) from electronic health records sources are representative of the national picture, while avoiding repeated work and replication in new IPD datasets. In our related publications on variation and trends in specific disease areas we report comparisons between trends in PCA data, and trends from other more labour-intensive sources such as CPRD, in more detail.

The UK government produces a 10-year trends document following the annual PCA data release, containing an overall summary of high-level trends and a brief breakdown of six interesting topics with the greatest level or change in prescribed items and cost [25]. However, the reported topics are few in number, chosen by NHS Digital, restricted to ten years of data, do not correct for inflation or population growth, are not easily discoverable by subject specialists, and readers are not able to interrogate their own topics of interest in detail. From 2016, the compiled datasets were also made available so users may conduct

their own exploration of the data, but, without drug names or categories being normalised, this is little better than the raw data, which we have processed into a normalised longitudinal dataset.

Policy implications and future research

Published papers can provide a useful and detailed insight into prescribing trends, but give a single snapshot which may quickly become out of date. Our tool facilitates ongoing monitoring by researchers and policymakers to assess prescribing changes in any area of concern or clinical interest they have identified; and permits interactive exploration of detailed issues in the data, such as individual presentations of chemicals, by any interested user. As part of our OpenPrescribing work we are using prescribing data to investigate adherence to guidelines and changes in practice in various clinical areas, to detect anomalous changes in individual practices relative to national trends in order to send practices alerts, and to identify cost saving opportunities. We have produced various manuscripts using the longitudinal data presented here as part of a range of data sources to describe variation in prescribing in primary care. We are happy to collaborate with other teams of clinicians and academics; we also release our underlying dataset and code as open data for re-use with citation.

Conclusions

Long-term trends in prescribing are interesting for a number of applications. While previous work on prescribing data has focused on static, manual analysis of a small number of drugs, modern data science approaches make it possible to create interactive services that allow clinicians, healthcare commissioners, policy makers, academics and any other interested party to interrogate and monitor prescribing trends for any combination of chemicals, to identify anomalies or signals of concern, and predict spending. We have delivered this using a combination of open data and freely accessible online tools.

ACKNOWLEDGEMENTS

We are grateful to Seb Bacon for maintaining databases and general assistance, and to Richard Croker for pharmaceutical advice and fuzzy matching in Excel.

FUNDING

OpenPrescribing is funded by: NIHR Oxford Biomedical Research Centre (BRC), Health Foundation (Ref 7599), NIHR SPCR (Ref 327). No specific funding was sought for this

project. The authors' funders had no involvement in the study design or the decision to submit.

CONFLICT OF INTEREST

All authors have completed the Unified Competing Interest form (available on request from the corresponding author) and declare: BG has received research funding from the Laura and John Arnold Foundation, the Wellcome Trust, the NHS NIHR School of Primary Care, the Health Foundation, NHS England, NIHR Oxford Biomedical Research Centre (BRC) and the WHO; he also receives personal income from speaking and writing for lay audiences on the misuse of science. HC is employed on BG's OpenPrescribing grants.

CONTRIBUTORSHIP STATEMENT

BG conceived and supervised the project, HC designed the methods, conducted the analysis, interpreted the findings, extracted and processed the data in BigQuery, Excel and Tableau with input from BG. HC and BG wrote the paper. All authors contributed to and approved the final manuscript. BG is guarantor.

DATA SHARING STATEMENT

The full compiled and processed PCA datasets are available online via Figshare [DOI 10.6084/m9.figshare.5447194.v1], SQL code is in Supplementary File, and the Trends tool (Tableau workbook) is available to use via https://OpenPrescribing.net/pca.

REFERENCE LIST

1 NHS-Digital. National Statistics Prescription Cost Analysis, England - 2016. 2016.http://www.content.digital.nhs.uk/searchcatalogue (accessed 28 Sep 2017).

2 NHS-Choices. About the National Health Service (NHS) in England - NHS Choices. 2016.http://www.nhs.uk (accessed 28 Sep 2017).

3 Mitchell AL, Hickey B, Hickey JL, *et al.* Trends in thyroid hormone prescribing and consumption in the UK. *BMC Public Health* 2009;**9**:132.

4 Gan EH, Pattman S, H S Pearce S, *et al.* A UK epidemic of testosterone prescribing, 2001-2010. *Clin Endocrinol* 2013;**79**:564–70.

5 Ilyas S, Moncrieff J. Trends in prescriptions and costs of drugs for mental disorders in England, 1998–2010. *Br J Psychiatry* 2012;**200**:393–8.

6 Dietrich ES. Effects of the National Institute for Health and Clinical Excellence's technology appraisals on prescribing and net ingredient costs of drugs in the National Health Service in England. *Int J Technol Assess Health Care* 2009;**25**:262–71.

7 Connor AJ, Fraser SG. Glaucoma prescribing trends in England 2000 to 2012. *Eye* 2014;**28**:863–9.

8 Fitzpatrick RW, Pate RG. Assessing the impact of NICE guidance on the prescribing of hormonal treatments of breast cancer in England. *J Eval Clin Pract* 2015;**21**:759–61.

9 NHS-Digital. Prescribing Data. 2017.https://digital.nhs.uk/article/4214/Prescribing (accessed 21 Sep 2017).

10 NHS-Digital. Prescription Cost Analysis - England, 2009 [NS]. 2012.https://digital.nhs.uk/catalogue/PUB01414 (accessed 11 Sep 2017).

11 NHS-Digital. Prescription Cost Analysis Glossary 2015. pres-cost-anal-eng-2015-apx.pdf. 2016.http://content.digital.nhs.uk/catalogue/PUB20200/pres-cost-anal-eng-2015-apx.pdf (accessed 28 Sep 2017).

12 Curtis HJ. Prescription Cost Analysis 1998-2016 - drugs matched to BNF via fuzzy lookup. *figshare* Published Online First: 28 September 2017. doi:10.6084/m9.figshare.5450323.v1

13 Patient. Name Changes of Medicines. 2015.https://patient.info/health/name-changes-of-medicines (accessed 20 Sep 2017).

14 Curtis HJ. Prescription Cost Analysis 1998-2016 data processing and normalisation. 2017.https://gist.github.com/HelenCEBM/192307b3c671a391f5ad6b44a3676880 (accessed 28 Sep 2017).

15 Office for National Statistics. Population Estimates. ONS Statistical bulletin. 2017.https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2016 (accessed 11 Sep 2017).

16 Office for National Statistics. CPI All Items Index. ONS Time Series Data. 2017.https://www.ons.gov.uk/economy/inflationandpriceindices/timeseries/d7bt/mm23 (accessed 20 Sep 2017).

17 RStudio-Inc. Shiny. 2017.https://shiny.rstudio.com/ (accessed 27 Sep 2017).

18 Bostock M. D3.js - Data-Driven Documents. 2017.https://d3js.org/ (accessed 27 Sep 2017).

19 Curtis HJ, Goldacre B. NHS prescription cost analysis data 1998-2016. *figshare* Published Online First: 28 September 2017. doi:10.6084/m9.figshare.5447194.v1

20 Bateman DN, Good AM, Afshari R, *et al.* Effects of licence change on prescribing and poisons enquiries for antipsychotic agents in England and Scotland. *Br J Clin Pharmacol* 2003;**55**:596–603.

21 Szatkowski L, Coleman T, McNeill A, *et al.* The impact of the introduction of smoke-free legislation on prescribing of stop-smoking medications in England. *Addiction* 2011;**106**:1827–34.

22 Langley TE, Huang Y, McNeill A, *et al.* Prescribing of smoking cessation medication in England since the introduction of varenicline. *Addiction* 2011;**106**:1319–24.

23 Szatkowski L, Aveyard P. Provision of smoking cessation support in UK primary care: impact of the 2012 QOF revision. *Br J Gen Pract* 2016;**66**:e10–5.

24 Acheampong P, Cooper G, Khazaeli B, *et al.* Effects of MHRA drug safety advice on time trends in prescribing volume and indices of clinical toxicity for quinine. *Br J Clin Pharmacol* 2013;**76**:973–9.

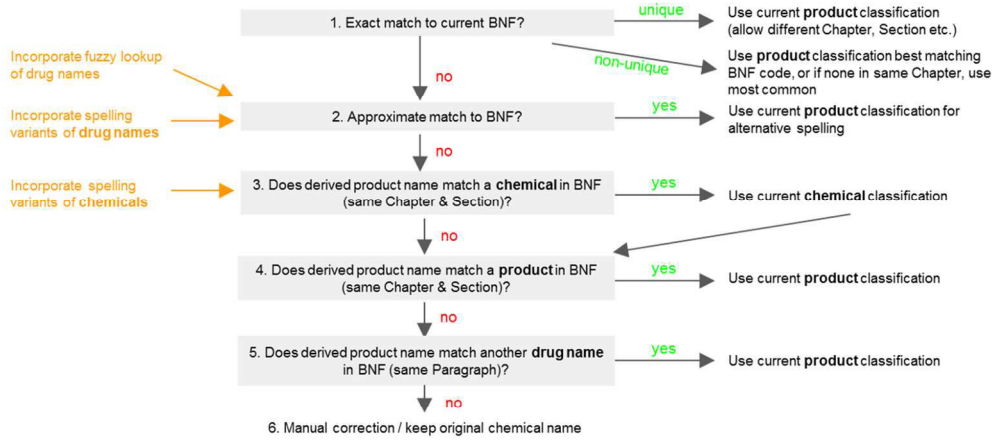
25 NHS-Digital. Prescriptions Dispensed in the Community, England 2006 to 2016. *Prescribing and Medicines Team, NHS Digital, UK* Published Online First: 29 June 2017.https://digital.nhs.uk/catalogue/PUB30014

LEGENDS TO FIGURES

Figure 1. *BNF code normalisation process flow chart describing how drug names were matched to the current BNF. Examples of matches at each numerated stage are given in table (bottom), with code/name changes in red. The first example demonstrates a drug which was matched to the current BNF through an exact match by name, but had moved from Chapter 4 to Chapter 2. For other types of matching, care was taken to avoid mistakenly matching to similar (but distinctly different) drugs across Chapters/Sections.*

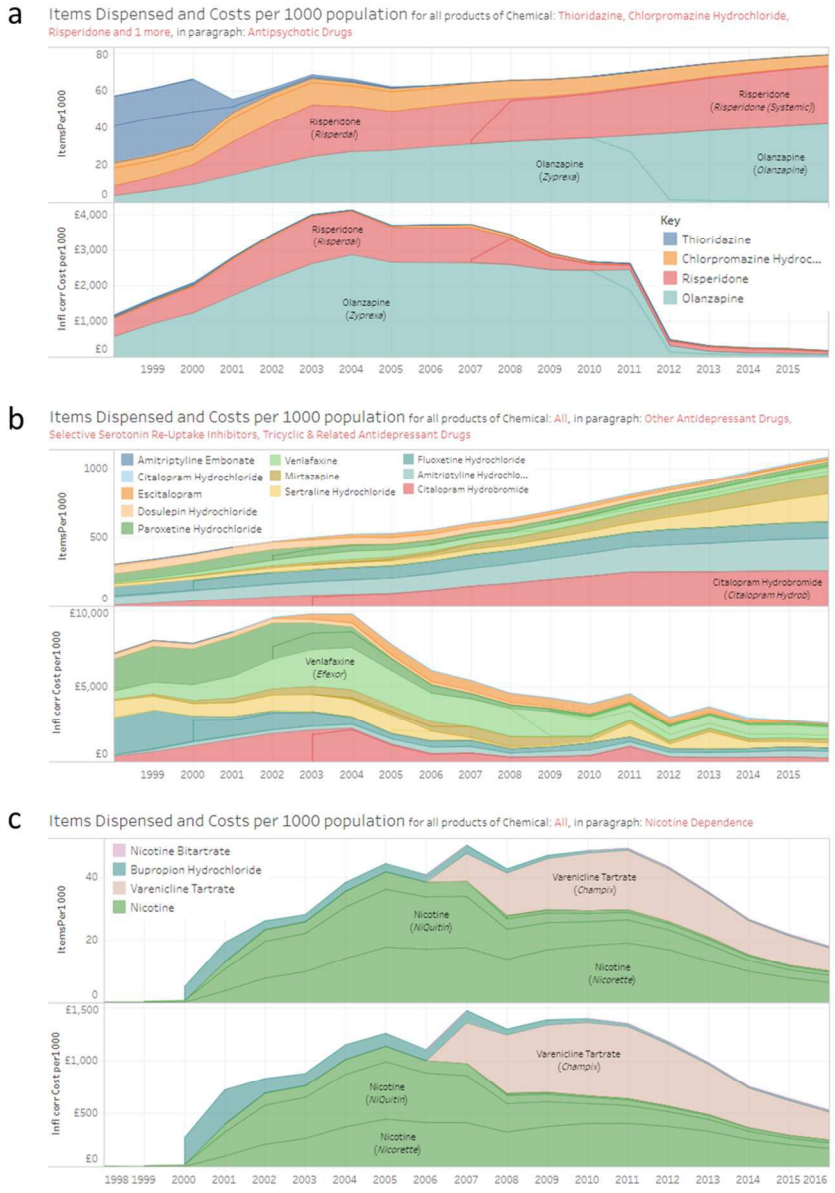
Figure 2. *Screenshots from Trends tool, showing items per 1,000 population and inflation-corrected costs per 1,000 population for selected drugs. Product names are in parentheses. (a) Prescribing trends for four selected antipsychotic chemicals following the safety alert on thioridazine [20]. Full dashboard available at https://public.tableau.com/shared/XX7DTWSG2?:display_count=yes. (b) Prescribing trends for selected antidepressant chemicals [5]. Full dashboard available at https://public.tableau.com/shared/72SJGGP89?:display_count=yes. (c) Prescribing trends for all chemicals within the Paragraph of Nicotine Dependence (smoking cessation medications). Full dashboard available at https://public.tableau.com/shared/6BW9J5RJB?:display_count=yes.*

Figure 3. *Screenshots from Trends tool, showing inflation-corrected costs per 1,000 population (a) by Chapter, and (b) by Section for Chapter 2 (Cardiovascular System).*

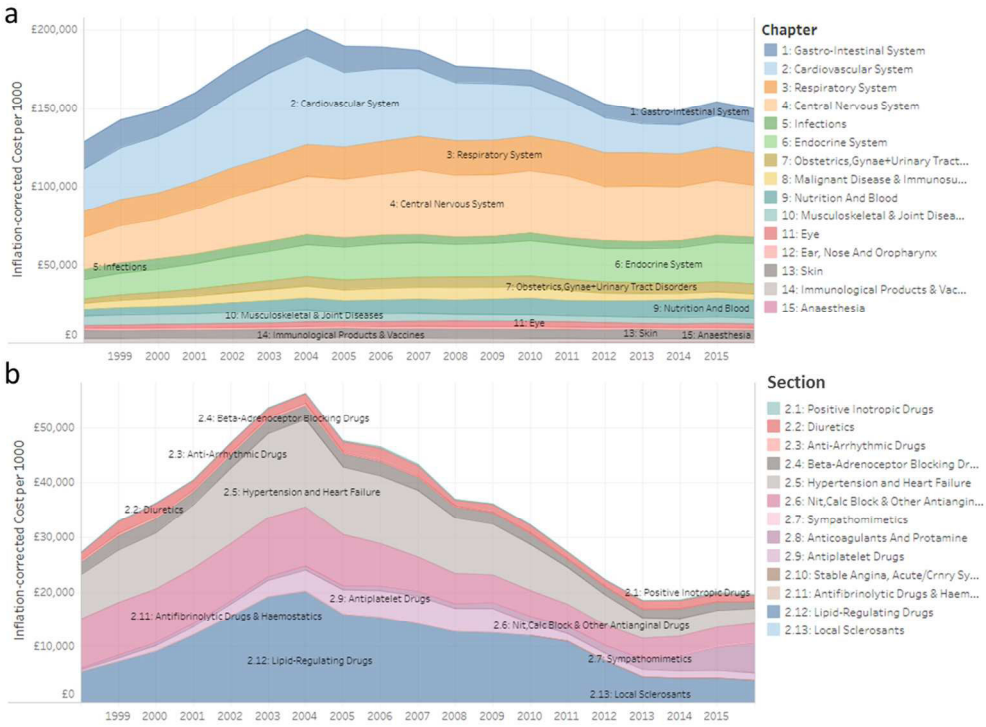


Match Type	Drug Name Original	Chemical Original	BNF Code Original	Drug Name Current	Chemical Current	Code Derived	Product Current
1	Aspirin_Pdr Sach 37.5mg	Aspirin	0407010	Aspirin_Pdr Sach 37.5mg	Aspirin	0209000A0AA	Aspirin (Antiplatelet)
2	Frusemide_Tab 20mg	Furosemide	0202020	Furosemide_Tab 20mg	Furosemide	0202020L0AA	Furosemide
2	Amlodipine Besyl_Liq Spec 10mg/5ml	Amlodipine Besylate	0206020	Amlodipine_Oral Soln 5mg/5ml	Amlodipine	0206020A0AA	Amlodipine
3	Felendil XL_Tab 10mg P/R	Felodipine	0206020	(none)	Felodipine	0206020F0	Felendil XL
4	Cozaar-Comp_Tab	Losartan Potassium With Diuretic	0205052	(none)	Losartan Potassium With Diuretic	0205052P0BB	Cozaar-Comp
5	Atorvastatin_Liq Spec 20mg/5ml	Atorvastatin	0212000	Atorvastatin_Oral Soln 20mg/5ml	Atorvastatin	0212000B0AA	Atorvastatin
6	Terbut Sulph_Syr 1.5mg/5ml S/F	Terbutaline Sulphate	0301011	(none)	Terbutaline Sulfate	0301011V0AA	Terbut Sulf

221x137mm (150 x 150 DPI)



219x311mm (150 x 150 DPI)



221x160mm (150 x 150 DPI)

SUPPLEMENTARY FILE

OpenPrescribing: Normalised Data and Software Tool to Research Trends in English NHS Primary Care Prescribing 1998-2016

Contents:

Table S1. Summary of original (unprocessed) PCA data by year (drugs in Chapters 1-15 only).

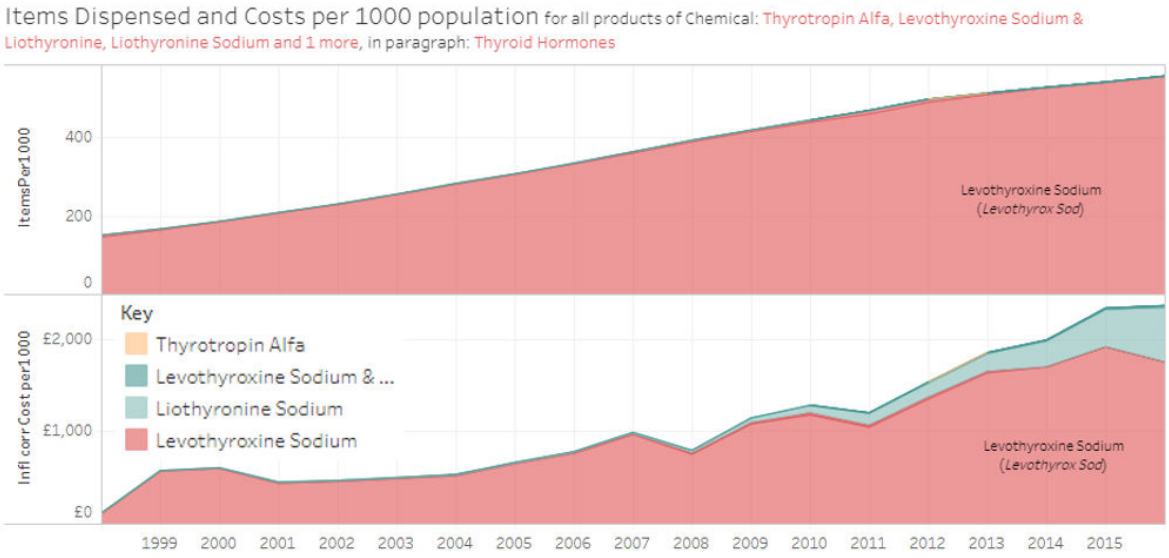
Year	Distinct count of Drug Name	Items	Cost
1998	6,338	497M	£4,440M
1999	6,587	513M	£5,011M
2000	6,613	535M	£5,284M
2001	6,754	569M	£5,784M
2002	6,834	599M	£6,487M
2003	6,893	630M	£7,113M
2004	6,912	666M	£7,645M
2005	6,907	699M	£7,452M
2006	6,810	728M	£7,660M
2007	7,056	772M	£7,810M
2008	7,202	817M	£7,716M
2009	7,401	859M	£7,892M
2010	11,703	898M	£8,162M
2011	11,751	932M	£8,101M
2012	12,207	969M	£7,802M
2013	12,318	996M	£7,846M
2014	12,576	1,027M	£8,022M
2015	12,875	1,043M	£8,403M
2016	13,285	1,061M	£8,284M
Total	***27,473***	**14,811M**	**£136,914M**

Table S2. Summary of drug code changes within the 1998-2016 prescribing datasets, by year.

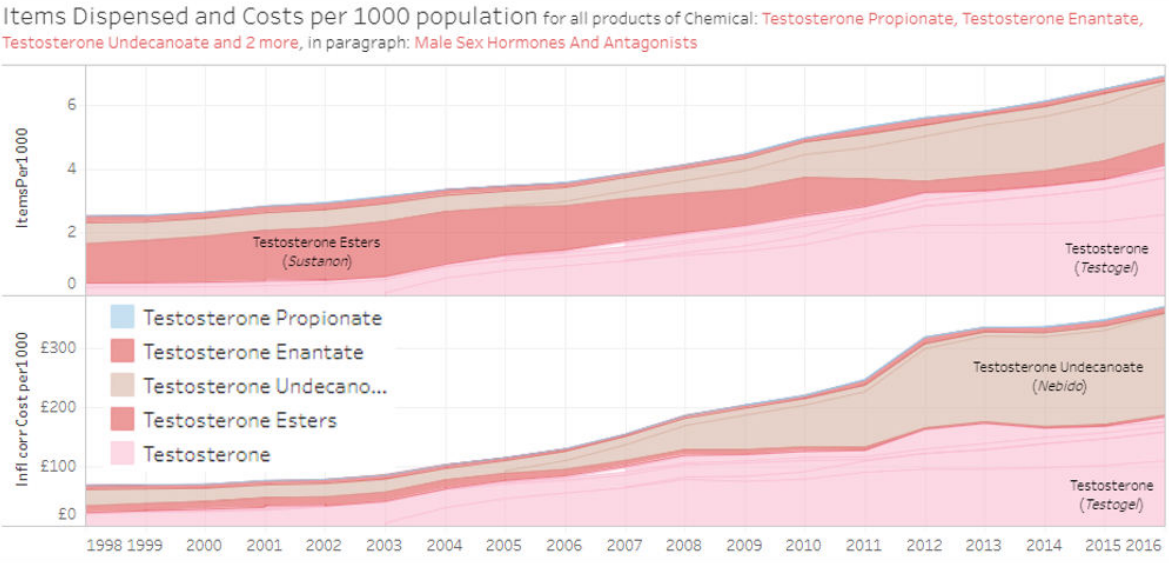
	Code change							
year	Chapter	Section	Paragraph	Sub-Paragraph	no change	no product match	no chemical match	Grand Total
1998	21	9	219	140	5,609	230	110	**6,338**
1999	21	7	188	88	5,917	237	129	**6,587**
2000	22	8	189	86	5,928	260	120	**6,613**
2001	26	9	178	85	6,069	278	111	**6,756**
2002	31	10	182	77	6,173	274	87	**6,834**
2003	34	9	191	79	6,256	241	83	**6,893**
2004	36	16	191	77	6,265	240	87	**6,912**
2005	26	14	157	85	6,315	245	65	**6,907**

2006	27	17	158	72	6,245	232	59	**6,810**
2007	35	12	168	74	6,483	237	47	**7,056**
2008	34	12	169	74	6,648	231	34	**7,202**
2009	3	4	180	76	6,919	219		**7,401**
2010	12	8	207	14	11,182	279	1	**11,703**
2011	15	4	130	14	11,331	257		**11,751**
2012	8	4	4	16	11,946	229		**12,207**
2013	2	4	8	1	12,139	163	1	**12,318**
2014		6	15	1	12,415	139		**12,576**
2015		6	6		12,770	93		**12,875**
2016		6	7		13,241	31		**13,285**

a


Items Dispensed and Costs per 1000 population for all products of Chemical: Thyrotropin Alfa, Levothyroxine Sodium & Liothyronine, Liothyronine Sodium and 1 more, in paragraph: Thyroid Hormones
ItemsPer1000
400
200
0
Levothyroxine Sodium
(Levothyrox Sod)
Infl corr Cost per1000
£2,000
£1,000
£0
Key
Thyrotropin Alfa
Levothyroxine Sodium & ...
Liothyronine Sodium
Levothyroxine Sodium
Levothyroxine Sodium
(Levothyrox Sod)
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

b


Items Dispensed and Costs per 1000 population for all products of Chemical: Testosterone Propionate, Testosterone Enantate, Testosterone Undecanoate and 2 more, in paragraph: Male Sex Hormones And Antagonists
ItemsPer1000
6
4
2
0
Testosterone Esters
(Sustanon)
Testosterone
(Testogel)
Infl corr Cost per1000
£300
£200
£100
£0
Testosterone Propionate
Testosterone Enantate
Testosterone Undecano...
Testosterone Esters
Testosterone
Testosterone Undecanoate
(Nebido)
Testosterone
(Testogel)
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

c

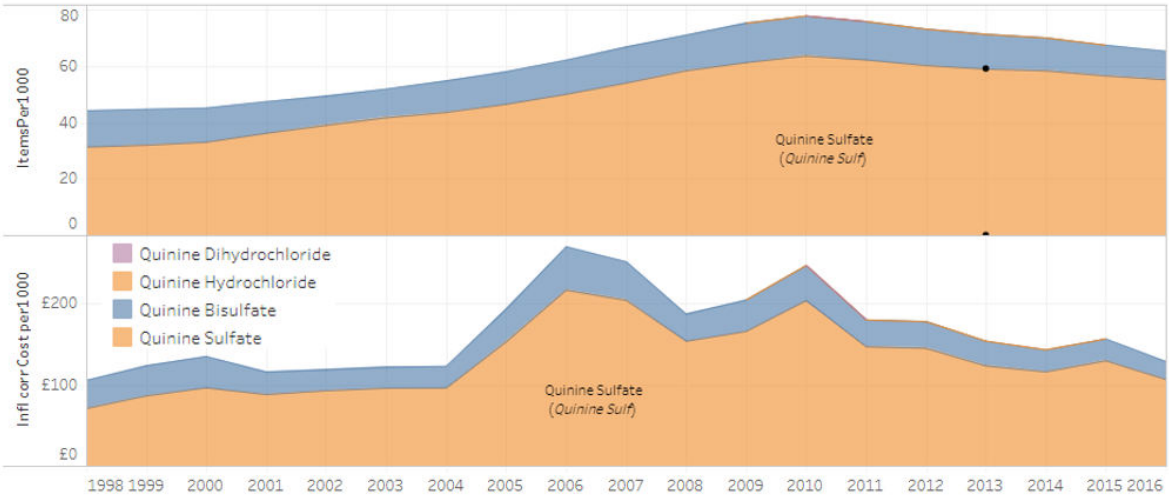

Items Dispensed and Costs per 1000 population for all products of Chemical: Quinine Dihydrochloride, Quinine Hydrochloride, Quinine Bisulfate and 1 more, in paragraph: Antimalarials
ItemsPer1000
80
60
40
20
0
Quinine Sulfate
(Quinine Sulf)
Infl corr Cost per1000
£200
£100
£0
Quinine Dihydrochloride
Quinine Hydrochloride
Quinine Bisulfate
Quinine Sulfate
Quinine Sulfate
(Quinine Sulf)
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

Figure S1. Screenshots from Trends tool, showing items per 1,000 population and inflation-corrected costs per 1,000 population for selected drugs. Product names are in parentheses. (a) Prescribing trends for all chemicals within the Paragraph of Thyroid Hormones. Full dashboard available at https://public.tableau.com/shared/GPW28PWJY?:display_count=yes. (b) Prescribing trends for all testosterone chemicals within the Paragraph of Male Sex Hormones. Full dashboard available at https://public.tableau.com/shared/YQ3ZFB3HY?:display_count=yes. c) Prescribing trends for all chemical forms of quinine (all of which are in the Antimalarials Paragraph). Full dashboard available at https://public.tableau.com/shared/85KJ2ZFN4?:display_count=yes.

Appendix - SQL Code for Processing and Normalisation of PCA data

A - Lookup Tables

A1. The special_cases lookup table

This is a workaround to assign a 'most likely' classification to the few problematic drug names which exist multiple times in BNF.

Lookup table is created by running the following script:

```
WITH temp as (
SELECT SUBSTR(SECTION_CODE,1,2) as chapter, section_code, presentation,
COUNT(DISTINCT product_code) as num
FROM ( SELECT DISTINCT section_code, section, para, subpara, chemical, product,
product_code, presentation FROM ebmdatalab.hscic.bnf )
GROUP BY chapter, section_code, presentation
HAVING num >1   --where name maps to more than one bnf code
ORDER BY chapter, num DESC)

SELECT
section_code, section, para, subpara, chemical, product, product_code,
presentation  -- this level is to filter to the top-prescribed code for each
drug name (according to latest detailed monthly data) (or, if none were
prescribed, then the first product name alphabetically
FROM (
  SELECT  -- this level joins all possible product codes to aggregated
prescribing data (2011-16) and ranks by items prescribed.
    a.*, b.items AS items_2011_2016,
    row_number() OVER (PARTITION BY a.presentation ORDER BY b.items DESC) AS
ranking -- We can use this to select the top/most likely drug code
  FROM (  -- this level is to look up all possible product codes for each drug
name in current BNF
      SELECT
```

```
      DISTINCT -- here we just want to go to product level rather than
individual presentations
      presentation,
      chapter, chapter_code,
      section, section_code,
      para, para_code,
      subpara, subpara_code,
      chemical, chemical_code,
      product, product_code
      from ebmdatalab.hscic.bnf where presentation in (select presentation from
temp where chapter < '18')

      ) a  -- now join to aggregated dataset grouped up to product level:
      LEFT JOIN ( SELECT substr(bnf_code,1,11) AS product_code, sum(items) as
items from ebmdatalab.aggregated_data.all_prescribed_BNFs_UpToSept2016 GROUP BY
product_code ) b
            ON a.product_code = substr(b.product_code,1,11)
      )
WHERE ranking = 1
ORDER BY chapter_code, presentation, product_code
```

A2. The lookup table of alternative drug spellings found within the data is created using the script below and saving as **ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_HC**

```
-- find drug name changes in PCA data to 2016
-- save results as ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_HC

WITH
a AS  (
  SELECT
      IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4)) AS
section_code, -- extra clause added to deal with those with extra spaces
2017-08-01

SUBSTR(drug_name,1,IF(STRPOS(drug_name,'_')>0,STRPOS(drug_name,'_')-1,length(dr
ug_name))) AS drug_name_part,  --take first part of drug name, up to underscore
(if there is one)
      MIN(year) AS min_year,  --this will help us to see which are the older vs
newer spellings used
      MAX(Year) AS max_year,
      SUM(items) AS Items
  FROM ebmdatalab.hscic.prescribing_pca_1998_2016_full
  WHERE  IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,2),SUBSTR(bnf_7_char,1,2))
< '18'
  GROUP BY
```

```
  section_code,
  drug_name_part
  ),

 b AS
(SELECT DISTINCT
section_code,
drug_name_part,
REPLACE(REPLACE(drug_name_part, 'i', '__'),'y','__') AS IY,
REPLACE(REPLACE(drug_name_part, 's', '__'),'z','__') AS SZ,
REPLACE(REPLACE(drug_name_part, 'ph', '__'),'f','__') AS PHF,
CONCAT(drug_name_part,'e') AS E,   -- add and E on to the end (note this only
works for the LAST word)
REPLACE(drug_name_part, ' ', ' e') AS E_mid -- add an E on to the end of all
words occurring before a space
FROM a)


SELECT
a.section_code,
a.drug_name_part,
CAST(a.min_year AS STRING) AS start_date,
CAST(a.max_year AS STRING) AS end_date,
a.items,
b.drug_name_part AS alternative,
CASE WHEN REPLACE(REPLACE(a.drug_name_part, 'i', '__'),'y','__') = b.IY THEN
'i_y'
      WHEN  REPLACE(REPLACE(a.drug_name_part, 's', '__'),'z','__') = b.SZ THEN
's_z'
      WHEN  REPLACE(REPLACE(a.drug_name_part, 'ph', '__'),'f','__') = b.PHF
THEN 'ph_f'
      WHEN  a.drug_name_part = b.E OR CONCAT(a.drug_name_part,'e') =
b.drug_name_part THEN 'e_end'
      WHEN  a.drug_name_part = b.E_mid OR REPLACE(a.drug_name_part, ' ', 'e ')
=b.drug_name_part THEN 'e_end'
      END AS type
FROM a
INNER JOIN b
  ON (REPLACE(REPLACE(a.drug_name_part, 'i', '__'),'y','__') = b.IY
      OR REPLACE(REPLACE(a.drug_name_part, 's', '__'),'z','__') = b.SZ
      OR REPLACE(REPLACE(a.drug_name_part, 'ph', '__'),'f','__') = b.PHF
      OR a.drug_name_part = b.E   --note this will only show this match once,
so we put in the other way around also
      OR a.drug_name_part = b.E_mid
      OR CONCAT(a.drug_name_part,'e') = b.drug_name_part
      OR REPLACE(a.drug_name_part, ' ', 'e ') =b.drug_name_part)
  AND a.drug_name_part != b.drug_name_part
  AND a.section_code = b.section_code
```

```
ORDER BY items desc
```

A3. The lookup table of Chemical name changes is created using the script below and saving as **pca_chemical_old_to_new_lookup**

```
-- PCA data - finding up to date chemical to combine with dataset
--save results as ebmdatalab.hscic.pca_chemical_old_to_new_lookup
WITH A as (
  SELECT
    IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4)) AS
Section,
    drug_name,
    count(distinct chemical) AS chems,
    max(year) AS Max_year_overall
    FROM
    ebmdatalab.hscic.prescribing_pca_1998_2016_full
    where
IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,2),SUBSTR(bnf_7_char,1,2)) <'18'
    GROUP BY
    Section,
    drug_name
    HAVING chems >1
    ),

B AS (
SELECT IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4))
AS section, drug_name, chemical,
    min(year) AS Min_year,
    max(year) AS Max_year
    FROM ebmdatalab.hscic.prescribing_pca_1998_2016_full
GROUP BY Section, drug_name, chemical
    ),

C AS (
SELECT DISTINCT
    A.Section,
    A.drug_name,
    B.chemical,
    B.min_year,
    B.max_year,
    IF(max_year = Max_year_overall,1,0) AS latest
    FROM A LEFT JOIN B ON A.drug_name = B.drug_name AND A.Section = B.Section
ORDER BY drug_name, chemical
)

SELECT old.section, old.drug_name, old.chemical AS old_chemical_name,
    nw.chemical AS new_chemical_name, nw.min_year AS Since
    FROM c old
```

```
     LEFT JOIN c nw ON old.drug_name = nw.drug_name AND old.chemical !=
nw.chemical and nw.latest = 1
    WHERE old.latest = 0
ORDER BY old.section, old.drug_name
```

A4. Known drug name changes

As reported online by patient.info
`ebmdatalab.hscic.drug_name_changes_2013`

A5. Fuzzy lookup for drugs not matching to BNF

List of drugs not matching BNF, identified through earlier iterations of the code.
These 1,084 drugs were matched to similar BNF names via fuzzy lookup in Excel and manually checked by a pharmacist.

List available at:
https://docs.google.com/spreadsheets/d/1UweKIZOLrKEzCtLULk5R5kJ4UyFIttvEouQ7RFGYrEA/edit#gid=594622641
and stored as `ebmdatalab.hscic.pca_bnf_name_to_code_fuzzy_lookup`

B - Data Extraction And Normalisation

B1. The latest chemical name for each drug is appended into the full dataset, to create prescribing_pca_1998_2016_full_v2

This does not take into account spelling changes but those will be handled later

```
-- save results as ebmdatalab.hscic.prescribing_pca_1998_2016_full_v2

SELECT a.*, COALESCE(c2.new_chemical_name, a.chemical) AS new_chemical_name
FROM
ebmdatalab.hscic.prescribing_pca_1998_2016_full a

LEFT JOIN ebmdatalab.hscic.pca_chemical_old_to_new_lookup c2
   ON a.drug_name = c2.drug_name
   AND a.chemical = c2.old_chemical_name
   AND IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4)) =
c2.section
```

C2. Run final data extraction parts 1 and 2 (scripts copied and updated from Issues #6 and #7)

B2a. Part 1

-- Final PCA data extraction part 1 (2016)

```
-- save results as ebmdatalab.tmp_eu.trends_from_pca

WITH
temp AS
(SELECT DISTINCT X.section_code, X.drug_name_part AS old_name, X.alternative
FROM
ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_HC X
INNER JOIN ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_2016_HC Y ON
X.alternative = Y.drug_name_part AND Y.end_date = '2016'
),

b AS (
  SELECT DISTINCT
   chapter_code, chapter, section_code, section, para, subpara, chemical,
product, product_code,
   REPLACE(presentation,'GlucOsamine','prop-GlucOsamine') AS
presentation,REPLACE(presentation,' ','')
     AS presentation_no_spaces
   FROM ebmdatalab.hscic.bnf
   WHERE presentation NOT IN (SELECT presentation from
ebmdatalab.hscic.bnf_name_to_product_special_cases_helen)
   AND chapter_code <'18'),

a0 AS (
SELECT *,
      TRIM(bnf_7_char) AS bnf_7_char_trim,

SUBSTR(drug_name,1,IF(STRPOS(drug_name,'_')>0,STRPOS(drug_name,'_')-1,length(dr
ug_name)))
        AS drug_name_part,
      SUBSTR(drug_name,1,IF(STRPOS(drug_name,' ')>0,STRPOS(drug_name,'
')-1,length(drug_name)))
        AS drug_name_part_short,
      SUBSTR(chemical,1,IF(STRPOS(chemical,' ')>0,STRPOS(chemical,'
')-1,length(chemical)))
        AS chemical_short,
      REPLACE(drug_name,'GlucOsamine','prop-GlucOsamine') AS drug_name_a,

REPLACE(REPLACE(drug_name,'GlucOsamine','prop-GlucOsamine'),'Sulph','Sulf') AS
drug_name_b,
      REPLACE(new_chemical_name,'Sulph','Sulf') AS new_chemical_name_b
```

```
      FROM ebmdatalab.hscic.prescribing_pca_1998_2016_full_v2 a
      WHERE
IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,2),SUBSTR(bnf_7_char,1,2)) < '18'),

a1 AS (SELECT a0.*,
        z.new_bnf_code AS code_fuzzy,
        z.new_name AS drug_name_fuzzy,
        CONCAT( UPPER(substr(d.new_name,1,1)),
substr(D.new_name,2,LENGTH(D.new_name)-1) ) AS product_2013,
        E.alternative AS product_new_spelling,
        CONCAT( UPPER(substr(d1.new_name,1,1)),
substr(D1.new_name,2,LENGTH(D1.new_name)-1) ) AS chemical_2013,  -- note, this
capitalises the first letter only
        replace(a0.new_chemical_name_b,a0.chemical_short,D3.new_name) AS
chemical_2013b,
        replace(a0.drug_name_b,a0.drug_name_part,e.alternative) AS
converted_drug_name, -- incorporate new spellings into drug name
        replace(a0.drug_name_b,a0.drug_name_part,D.new_name) AS
converted_drug_name2,
        replace(a0.drug_name_b,a0.drug_name_part_short,D2.new_name) AS
converted_drug_name3,

SUBSTR(drug_name_b,1,IF(STRPOS(drug_name_b,'_')>0,STRPOS(drug_name_b,'_')-1,len
gth(drug_name_b))) AS drug_name_part_b

        FROM a0
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D ON
LOWER(drug_name_part) = D.old_name
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D1 ON
LOWER(a0.chemical) = D1.old_name
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D2 ON
LOWER(drug_name_part_short) = D2.old_name
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D3 ON
LOWER(a0.chemical_short) = D3.old_name
      LEFT JOIN temp E ON a0.drug_name_part = E.old_name AND
SUBSTR(a0.bnf_7_char_trim,1,4) = e.section_code
      LEFT JOIN ebmdatalab.hscic.pca_bnf_name_to_code_fuzzy_lookup z ON
A0.drug_name = z.old_name
      ),


--CAPITALISE WHERE NEEDED:
A2 AS (
SELECT *,
      CONCAT( UPPER(substr(converted_drug_name2,1,1)),
substr(converted_drug_name2,2,LENGTH(converted_drug_name2)-1) ) AS
converted_drug_name4,
```

```
      CONCAT( UPPER(substr(converted_drug_name3,1,1)),
substr(converted_drug_name3,2,LENGTH(converted_drug_name3)-1) ) AS
converted_drug_name5,
      CONCAT( UPPER(substr(chemical_2013b,1,1)),
substr(chemical_2013b,2,LENGTH(chemical_2013b)-1) ) AS chemical_2013_c
--capitalise chemical as well.
      FROM A1
      ),

--COALESCE TO FORM "FINAL" NAMES
a3 AS (
 SELECT *,
COALESCE(converted_drug_name,converted_drug_name4,converted_drug_name5,drug_nam
e_fuzzy,drug_name_b)
   AS drug_name_F,
 COALESCE(product_new_spelling,product_2013,drug_name_part_b) AS
drug_name_part_F,
 COALESCE(chemical_2013, chemical_2013_c, new_chemical_name_b) AS chemical_F
 from A2
 ),

 --add a drug name field without spaces:
 a4 AS (
 select *, REPLACE(drug_name_F,' ','')  as drug_name_F_no_spaces
 from A3
 ),


a AS  (
  SELECT
      x.bnf_7_char_trim AS bnf_code,
      x.drug_name,
      drug_name_F,
      COALESCE(spc.presentation,b.presentation,ba.presentation,
bb.presentation,bc.presentation, bd.presentation)
        AS current_bnf_name,
      COALESCE(spc.product_code,b.product_code,ba.product_code,
bb.product_code,bc.product_code, bd.product_code)
        AS current_bnf_code,
      drug_name_part,
      drug_name_part_F, -- use as product name if no other
      x.section,
      x.subpara,
      x.chemical AS Chemical_original,
      x.chemical_F AS Chemical,
      x.Year,
      SUM(x.owc2) AS OWC2,   -- prescribed generically but no generic available
      SUM(x.NIC) AS Cost,
```

```
      SUM(x.items) AS Items,
      SUM(x.quantity) AS Quantity
  FROM a4 x
    --AND A.Currently_in_BNF = 'N'
    LEFT JOIN ebmdatalab.hscic.bnf_name_to_product_special_cases_helen spc ON
upper(x.drug_name_F) = upper(spc.presentation)  -- look up original drug
details in current bnf (drugs matching more than one drug in bnf)
    LEFT JOIN b ON upper(x.drug_name_F) = upper(b.presentation)   -- use upper
to match up examples like this: "Pentasa Sr_Tab 250mg" and "Pentasa SR_Tab
250mg"
            AND SUBSTR(x.bnf_7_char_trim,1,4) = b.section_code   -- look up
original drug details in current bnf.

    LEFT JOIN b ba ON upper(x.drug_name_F) = upper(ba.presentation)
            AND SUBSTR(x.bnf_7_char_trim,1,4) != ba.section_code  -- check if
drug now only belongs in a different section but same chapter
            AND SUBSTR(x.bnf_7_char_trim,1,2) = ba.chapter_code
    LEFT JOIN b bb ON upper(x.drug_name_F) = upper(bb.presentation)
            AND SUBSTR(x.bnf_7_char_trim,1,2) != bb.chapter_code  -- check if
drug now only belongs in a different chapter
    LEFT JOIN b bc ON upper(x.drug_name) = upper(bc.presentation) AND
b.presentation IS NULL AND ba.presentation IS NULL AND bb.presentation IS NULL
--also check original in case new drug name didn't work e.g. nifedipin(e)
            AND SUBSTR(x.bnf_7_char_trim,1,4) = bc.section_code
    LEFT JOIN b bd ON x.drug_name_F_no_spaces = bd.presentation_no_spaces   --
match without spaces e.g. Terbut Sulf_Inha 250mcg (400 D) vs "(400D)"
            AND SUBSTR(x.bnf_7_char_trim,1,4) = bd.section_code  AND
b.presentation IS NULL AND ba.presentation IS NULL AND bb.presentation IS NULL
-- look up original drug details in current bnf.

  GROUP BY
      bnf_code, drug_name, drug_name_F, current_bnf_name, current_bnf_code,
      drug_name_part, drug_name_part_F, -- use as product name if no other
      section, subpara, Chemical_original, Chemical, Year
      )

SELECT
    a.bnf_code,
    a.current_bnf_code AS Product_code_updated,
    SUBSTR(COALESCE(a.current_bnf_code,b.product_code,a.bnf_code),1,2) AS
Chapter_code_current,
    SUBSTR(a.bnf_code,1,2) AS BNF_Chap_Code,
    COALESCE(b.chapter, ch.description) AS Chapter_Current,
    ch.description AS Chapter_original,
    SUBSTR(COALESCE(a.current_bnf_code,a.bnf_code),3,2) AS
Section_code_current,
    SUBSTR(bnf_code,3,2) AS BNF_Section_Code,
    COALESCE(b.section,se.description,a.section) AS Section_Current,
```

```
    a.section AS Section_Original,
    SUBSTR(COALESCE(a.current_bnf_code,b.product_code,a.bnf_code),5,2) AS
Para_code_current,
    COALESCE(b.para,pa.description) As Para_current,
    COALESCE(b.subpara,a.subpara) As Subpara_current,
    a.subpara AS Subpara_original,
    COALESCE(b.chemical,a.chemical) As Chemical_current,
    a.Chemical_original,
    COALESCE(b.product, a.drug_name_part_F) AS Product_current,
    current_bnf_name,
    a.drug_name,
    IF(b.product_code IS NULL,'N','Y') AS Currently_in_BNF,
    a.year,
    a.Items,
    a.owc2,
    a.Quantity,
    a.Cost

  FROM a
  LEFT JOIN ebmdatalab.hscic.bnf_vertical ch ON SUBSTR(a.bnf_code,1,2) =
ch.code
  LEFT JOIN ebmdatalab.hscic.bnf_vertical se ON SUBSTR(a.bnf_code,1,4) =
se.code
  LEFT JOIN ebmdatalab.hscic.bnf_vertical pa ON SUBSTR(a.bnf_code,1,6) =
pa.code
  LEFT JOIN b ON a.current_bnf_name = b.presentation
            AND a.current_bnf_code = b.product_code
Save results as ebmdatalab.tmp_eu.trends_from_pca_2016
```

B2b. Part 2

```
-- final pca data extraction (2016) part 2
-- distinct product-chemical combinations in current BNF:
WITH
 chem_p AS (
   SELECT DISTINCT product, product_code, chemical_code, chemical,
   count (distinct product_code) Over (partition by chemical_code, product)
      AS Dist_prods_with_same_name
   FROM ebmdatalab.hscic.bnf
   WHERE chapter_code <'18'
   ORDER BY Dist_prods_with_same_name, product),

-- find all drug_name_parts in PCA which have been mapped to a new chemical:
chem_0 AS (
  SELECT
```

```
SUBSTR(drug_name,1,IF(STRPOS(drug_name,'_')>0,STRPOS(drug_name,'_')-1,length(dr
ug_name)))
      AS drug_name_part,
  drug_name, section, old_chemical_name, new_chemical_name
  FROM ebmdatalab.hscic.pca_chemical_old_to_new_lookup_2016),

-- distinct *chemicals* in current BNF:
chem_a AS (
SELECT chemical,
    count(distinct chapter) AS Chapters,
    count(distinct section) AS Sections,
    count(distinct para) AS Paras,
    count(distinct chemical_code) AS Codes,
    min(chemical_code) AS min_code
  FROM ebmdatalab.hscic.bnf
  WHERE chapter_code < '18'
  GROUP BY chemical
  ORDER BY codes DESC, paras DESC, chemical),

-- for chemicals with multiple codes:
-- check whether each chemical code is the only one in its paragraph / section
/ chapter
chem_a1 AS
(SELECT DISTINCT
  a.chemical, b.chemical_code, a.paras, b.para_code, a.sections,b.section_code,
a.chapters, b.chapter_code,
  count(distinct b.chemical_code) over (partition by b.chemical,chapter_code)
    AS appearances_by_chapter,
  count(distinct b.chemical_code) over (partition by b.chemical,section_code)
    AS appearances_by_section,
  count(distinct b.chemical_code) over (partition by b.chemical,para_code)
    AS appearances_by_para
  FROM ebmdatalab.hscic.bnf b
  INNER JOIN chem_a a ON a.chemical = b.chemical and a.codes > 1
  WHERE b.chapter_code < '18'
  ORDER BY chemical ),


-- SELECT ALL CHEMICALS FROM BNF WHICH MAP TO A SINGLE PRODUCT
-- used in final step only
b AS (
  SELECT DISTINCT
   chapter_code, chapter, section_code, section, para_code, para, subpara_code,
subpara, chemical_code
   FROM ebmdatalab.hscic.bnf
   WHERE chapter_code <'18'),
```

```
t as (
 SELECT t.*,
 REPLACE(c2.new_chemical_name,'Streptokinase-Streptodornase','Streptokinase &
Streptodornase')
      AS new_chemical_name,
 CASE WHEN Product_current LIKE 'Levonelle%' THEN '0703050A0BC'   --
'Levonelle'
      WHEN Product_current LIKE 'Postinor%' THEN '0703050A0BB'    -- 'Postinor'
      WHEN t.drug_name LIKE 'Terbut%Sulph_Syr%' THEN '0301011V0AA'  -- 'Terbut
Sulf'
      WHEN t.drug_name LIKE 'Thalidomide%' AND Chapter_code_current = '05' THEN
'0501100J0AA' -- 'Thalidomide (Antileprotic)'
      WHEN Product_current LIKE 'Menoring 50' THEN '0702010G0BE'  -- 'Menoring
50'
      WHEN t.drug_name = 'Acetylcy_Eye Dps 10% (Old)' THEN '1108010C0AA' --
'Acetylcy (Eye)'
      WHEN t.drug_name = 'Abilify Maintena_Inj 400mg Vl + Dil' THEN
'0402020ADBB' -- 'Abilify Maintena'
      WHEN Product_current LIKE 'Melatonin%' THEN '0401010ADAA' -- 'Melatonin'
      WHEN Product_current LIKE 'Varidase%' THEN '1311070R0BB' -- 'Varidase'
      WHEN t.drug_name = 'Cocois_Scalp Oint' THEN '1305020V0BB' -- 'Cocois'
      WHEN t.drug_name = 'Levocarnitine_Oral Soln Paed 1.5g/5ml30%' THEN
'0908010C0AA' -- 'Levocarnitine'
      ELSE product_code_updated
      END AS product_code_updated_manual,
 CASE WHEN Product_current LIKE 'Levonelle%' THEN 'Levonelle'
      WHEN Product_current LIKE 'Postinor%' THEN 'Postinor'
      WHEN t.drug_name LIKE 'Terbut%Sulph_Syr%' THEN 'Terbut Sulf'
      WHEN t.drug_name LIKE 'Thalidomide%' AND Chapter_code_current = '05' THEN
'Thalidomide (Antileprotic)'
      WHEN Product_current LIKE 'Menoring 50' THEN 'Menoring 50'
      WHEN t.drug_name = 'Acetylcy_Eye Dps 10% (Old)' THEN 'Acetylcy (Eye)'
      WHEN t.drug_name = 'Abilify Maintena_Inj 400mg Vl + Dil' THEN 'Abilify
Maintena'
      WHEN Product_current LIKE 'Melatonin%' THEN 'Melatonin'
      WHEN Product_current LIKE 'Varidase%' THEN 'Varidase'
      WHEN t.drug_name = 'Cocois_Scalp Oint' THEN 'Cocois'
      WHEN t.drug_name = 'Levocarnitine_Oral Soln Paed 1.5g/5ml30%' THEN
'Levocarnitine'
      ELSE product_current
      END AS product_current_manual
 FROM ebmdatalab.helen.trends_from_pca_2016 t
 LEFT JOIN chem_0 c2 ON t.drug_name = c2.drug_name AND t.chemical_current =
c2.old_chemical_name AND SUBSTR(t.bnf_code,1,4) = c2.section
 ),
```

```
  A AS (
 SELECT T.*,
 COALESCE(chem_p.product,c2.product,product_current_manual) AS current_product,
    -- use this order in coalesce because we want to update/replace any
existing product names for which we now have a better one.
 COALESCE(product_code_updated_manual,chem_p.product_code,c2.product_code)
    AS current_product_code,

COALESCE(chem_p.chemical,c2.chemical,chem_a.chemical,chem_a1.chemical,c3.chemic
al)
    AS unique_chem,  -- chemicals currently in BNF (uniquely)

COALESCE(chem_p.chemical_code,c2.chemical_code,chem_a.min_code,chem_a1.chemical
_code,c3.min_code)
    AS unique_chem_code

 FROM  t
  -- link to BNF using whole Product name (note this will be drug_name_part)
--------------------
  -- chemical must match as well because product names are not always unique.
  LEFT JOIN chem_p ON t.product_current = chem_p.product
            AND t.Product_code_updated_manual IS NULL
            AND (UPPER(chem_p.chemical) = UPPER(Chemical_current)
                      OR  UPPER(chem_p.chemical) = UPPER(new_chemical_name))
            AND SUBSTR(chem_p.chemical_code,1,6) = SUBSTR(bnf_code,1,6)
                      -- some chemicals sit in multiple paras.
            AND chem_p.Dist_prods_with_same_name = 1

  -- try shortening Product names in BNF to match products in data (only if
whole name is not found) --
    LEFT JOIN chem_p c2 ON t.product_current =
SUBSTR(c2.product,1,length(t.product_current))
            AND t.Product_code_updated_manual IS NULL
            AND chem_p.product IS NULL
            AND UPPER(c2.chemical) IN (UPPER(Chemical_current),
UPPER(new_chemical_name))
            AND SUBSTR(c2.chemical_code,1,6) = SUBSTR(bnf_code,1,6)  --some
chems sit in multiple paras.
            AND chem_p.Dist_prods_with_same_name = 1

  -- link to BNF using "original" chemical name for chemicals which are unique
in BNF  ---------------
  LEFT JOIN chem_a ON UPPER(chem_a.chemical)  = UPPER(Chemical_current)
            AND chem_a.codes = 1 AND chem_p.chemical IS NULL
            AND t.Product_code_updated_manual IS NULL
  -- link to BNF using NEW chemical name for chemicals which are unique in BNF
--------------------
  LEFT JOIN chem_a c3 ON c3.chemical  = new_chemical_name
```

```
            AND c3.codes = 1 AND chem_a.chemical IS NULL
            AND chem_p.chemical IS NULL
            AND t.Product_code_updated_manual IS NULL
  -- link to BNF using NEW chemical name for chemicals which are NON-unique in
BNF  --------------
  -- provided that no chemical has been assigned in a previous join.
  -- first check same paragraph then section then chapter.
  LEFT JOIN chem_a1 ON chem_a1.chemical  = Chemical_current
            AND chem_p.chemical IS NULL
            AND chem_a.chemical IS NULL
            AND c3.chemical IS NULL
            AND t.Product_code_updated_manual IS NULL
            AND (
              (chem_a1.para_code = SUBSTR(bnf_code,1,6) AND
chem_a1.appearances_by_para = 1)
              OR (chem_a1.section_code = SUBSTR(bnf_code,1,4) AND
chem_a1.appearances_by_section = 1)
              OR (chem_a1.chapter_code = SUBSTR(bnf_code,1,2) AND
chem_a1.appearances_by_chapter = 1)
                    )
 ORDER BY drug_name,year ),

 u AS (
 select bnf_code,
 Chapter_code_current,   BNF_Chap_Code,     Chapter_Current, Chapter_original,
 Section_code_current,   BNF_Section_Code,  Section_Current,   Section_Original,
 Para_code_current,      Para_current,
 Subpara_current,  Subpara_original,
 COALESCE(unique_chem_code, SUBSTR(Product_code_updated,1,9),
SUBSTR(current_Product_code,1,9))
   AS chem_code_today, --chemical code
 Chemical_original,
 COALESCE(unique_chem,Chemical_current) AS chem_today,
 COALESCE(Product_code_updated, current_product_code)    AS prod_code_today,
 COALESCE(current_product,     Product_current) AS prod_today,
   -- note this is opposite way around to code because we want to replace the
previous name
   -- but there may not be a code.
 current_bnf_name, drug_name,
 Currently_in_BNF,
 year,  Items,  owc2,  Quantity,  Cost
 FROM a
 ORDER BY drug_name, year)


SELECT
    bnf_code,
    COALESCE(b.chapter_code,Chapter_code_current) AS Chapter_code_current,
```

```
    BNF_Chap_Code,
    COALESCE(b.chapter,Chapter_Current) AS Chapter_Current,
    Chapter_original,
    COALESCE(SUBSTR(b.section_code,3,2),Section_code_current) AS
Section_code_current,
    BNF_Section_Code,
    COALESCE(b.section,Section_Current) AS Section_current,
    Section_Original,
    COALESCE(SUBSTR(b.para_code,5,2),Para_code_current) AS Para_code_current,
    COALESCE(b.para,Para_current) AS Para_current,
    COALESCE(b.subpara,Subpara_current) AS Subpara_current,
    Subpara_original,
    chem_code_today AS Chemical_code_current,
    Chemical_original,
    chem_today AS Chemical_current,
    prod_code_today AS Prod_code_current,
    prod_today AS product_current,
    current_bnf_name, drug_name,
    Currently_in_BNF, u.year,  Items,  owc2,  Quantity,  Cost,
    -- add calculated fields:
    1000*items/pop.Population AS ItemsPer1000,
    1000*quantity/pop.Population AS QuantityPer1000,
    Inf.Multiplier_2016*cost AS Infl_corr_Cost,
    1000*Inf.Multiplier_2016*cost/pop.Population AS Infl_corr_Cost_per1000,
    IEEE_DIVIDE(Inf.Multiplier_2016*Cost, Items) AS Infl_corr_CostPerItem,
    1000*owc2/pop.Population AS Owc2Per1000

 FROM U
 LEFT JOIN b ON u.chem_code_today = b.chemical_code
 LEFT JOIN ebmdatalab.ONS.england_midyear_population pop ON u.Year = pop.Year
 LEFT JOIN ebmdatalab.ONS.inflation_cpi inf ON u.Year = inf.Year
 --WHERE LENGTH(chem_code_today) =8
```

BMJ Open

OpenPrescribing: Normalised Data and Software Tool to Research Trends in English NHS Primary Care Prescribing 1998-2016

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019921.R1
Article Type:	Research
Date Submitted by the Author:	21-Nov-2017
Complete List of Authors:	Curtis, Helen; University of Oxford Department of Primary Care Health Sciences Goldacre, Ben; University of Oxford, Primary Care Health Sciences
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	General practice / Family practice
Keywords:	Prescribing, National Health Service (NHS), PRIMARY CARE

OpenPrescribing: Normalised Data and Software Tool to Research Trends in English NHS Primary Care Prescribing 1998-2016

Helen J Curtis[1], Ben Goldacre[1,*]

*Corresponding author: ben.goldacre@phc.ox.ac.uk

[1]Evidence Based Medicine DataLab
Centre for Evidence Based Medicine
Department of Primary Care Health Sciences
University of Oxford
Radcliffe Observatory Quarter
Woodstock Road
Oxford OX2 6GG

Word Count: 3,915

ABSTRACT

Objectives: We aimed to compile and normalise England's national prescribing data for 1998-2016 to facilitate research on long-term time trends, and create an open data exploration tool for wider use.

Design: We compiled data from each individual year's national statistical publications and normalised them by mapping each drug to its current classification within the national formulary where possible. We created a freely accessible, interactive web tool to allow anyone to interact with the processed data.

Setting and Participants: We downloaded all available annual prescription cost analysis datasets, which include cost and quantity for all prescription items dispensed in the community in England. Medical devices and appliances were excluded.

Primary and secondary outcome measures: We measured the extent of normalisation of data and aimed to produce a functioning accessible analysis tool.

Results: All data were imported successfully. 87.5% of drugs were matched exactly on name to the current formulary, and a further 6.5% to similar drug names. All drugs in core clinical Chapters were reconciled to their current location in the data schema, with only 1.26% of drugs not assigned a current chemical code. We created an openly accessible interactive tool to facilitate wider use of these data.

Conclusions: Publicly available data can be made accessible through interactive online tools, to help researchers and policymakers explore time trends in prescribing.

Strengths and limitations of this study

- We processed publicly-available annual data for the whole of England's community dispensing - not a sample.
- We corrected for population size, inflation, and (where possible) drugs changing name and/or classification over time.
- We produced a free, openly accessible tool for wider use, displaying trends in items, cost, price-per-item and quantity-per-item for each product for 1998-2016, which can be updated annually.
- The tool is limited to product-level data, not individual presentations, and wide-scale correction for dosage was not possible.
- Users can also download our normalised dataset in order to carry out their own analyses.

Abbreviations

BNF - British National Formulary
BSA - NHS Business Services Authority
CNS - Central nervous system
CPI - Consumer Price Index
CPRD - Clinical Practice Research Datalink
GIS - Gastrointestinal system
NHS - National Health Service
NIC - Net Ingredient Cost
ONS - Office of National Statistics
PCA - Prescription Cost Analysis

INTRODUCTION

In 2016, NHS prescribing in England cost £9.20bn [1], approximately 9% of the annual NHS budget [2]. Prescribing behaviour is expected to respond within the dynamic system of evidence-based medicine, through changing patterns of disease, innovation in medical treatments, and new evidence. Monitoring long-term time trends in prescribing is therefore useful to observe changes in practice, to provide a form of feedback to ensure there are no unexpected or undesirable changes, and to facilitate tracking and forecasting of costs.

NHS Digital publish monthly and annual prescribing datasets from the NHS Business Services Authority (NHSBSA), along with static reports on prescribing trends. However this does not allow readers to interrogate topics of interest in detail, and the large datasets can be complex to manage. We provide a service at OpenPrescribing that facilitates exploration of outliers and trends for individual general practices in NHS England, which has provided over 250,000 analyses to 50,000 users over the past year. The detailed dataset that drives this service (running to over 10m rows a month) is only available from 2010 onwards. The annual Prescription Cost Analysis (PCA) data, aggregated nationally (with no data on individual practice), and by year (with no data on prescribing changes each month) are available back to 1998. These data are freely accessible, but consist of individual files for each year of prescribing, which cannot be straightforwardly combined, and therefore do not facilitate interrogation of time trends. Additionally, identifiers for individual drugs may change name, or location within the British National Formulary (BNF), over time, making simple compilation of the data impossible.

The value of PCA data is indicated by the numerous previous studies using it to assess prescribing trends [3–5] or to detect changes in response to guidelines or safety alerts [6–8]. These studies have been focused on data for a small number of drugs, manually aggregated for each bespoke analysis; furthermore, given publishing delays for academic manuscripts the data are commonly very delayed, and readers cannot easily place the findings in context of current clinical practice or expenditure.

We therefore set out to aggregate all available PCA data into a single data frame for longitudinal analysis of trends, in a service that could be easily updated; to generate an interactive online service where any user can explore and monitor time trends in prescribing using the latest available data; and to share all resources for re-use by others as open data.

METHODS

Data sources

Every available Prescription Cost Analysis (PCA) annual dataset was downloaded from NHS Digital or National Archives, covering calendar years 1998 to 2016 [9].

Data structure

Each annual PCA dataset includes all items dispensed in England by pharmacy/appliance contractors, dispensing doctors, and items personally administered by doctors, whether or not they were *prescribed* in England or other parts of the UK. Items dispensed in other settings (prisons, hospitals and private prescriptions) are excluded. Prior to 2010, the data were rounded to the nearest 100 and excluded drugs with fewer than 50 items prescribed, accounting for 0.01% of total items [10]. Definitions of key terms used in the PCA data (and NHS primary care prescribing data more generally) are given in Box 1 and a full glossary of terms is available [11].

Box 1. Glossary of prescribing data terminology.

Example presentation: Tradorec XL Tablets 300mg

Drug Name	BNF Chemical Name	BNF Section Name	BNF Sub Paragraph Name	Items	Quantity	NIC (£)
Tradorec XL_Tab 300mg	Tramadol Hydrochloride	Analgesics	Opioid Analgesics	6,374	324,167	152,358

- The *drug name* describes the full *presentation* of the drug, i.e. the formulation and strength as well as the drug's brand or generic (*product*) name.
- The *chemical name* is the standard registered name for the active constituent of the medicine. It is not always an individual chemical: examples include "Paracetamol Combined Preparations" and "Paracetamol & Caffeine".
- Numerical codes representing Chapter, Section, Paragraph and Sub-paragraph are also supplied. These represent only the first seven characters of each drug's unique 15-character BNF code - see Box 2.
- *Items* are functionally equivalent to prescriptions; they do not take into account the quantity (number of boxes/bottles etc.) dispensed to the same person. *Items* may vary in the *quantity* prescribed.
- *Quantity* represents the quantity of a drug dispensed, with units of measurement (units/tablets/grammes/millilitres etc.) dependent upon its formulation.
- *Net Ingredient Cost* (NIC) represents the basic price of the medicine, i.e. the Drug Tariff price, or, if not listed, the price published by the manufacturer or supplier. NIC may be subject to further charges and/or discounts. Patients who are eligible contribute a fixed fee towards each prescription charge, but this only applies to a minority of items and it is not possible to identify which items in this dataset.

Every drug presentation (i.e. each formulation, dose and product combination) is described by a unique drug name, and has a unique 15-digit structured British National Formulary (BNF) code, an example of which is given in Box 2. The BNF contains an entry for every product available to be prescribed in Britain, including medicinal products, dietary supplements, complementary therapies and physical appliances such as bandages. The hierarchical BNF codes imply a data schema as follows: each *presentation* of a drug has a *product* name, which may be either a brand name or the generic *chemical* name; as such, each product can be mapped to a chemical. Each chemical is a member of a Paragraph in the BNF (some of which are divided into Sub-paragraphs, which themselves often approximate to a class of drugs). Each Paragraph belongs to a Section, which is in turn a member of a Chapter (often approximating to a system of the body, such as "Cardiovascular").

Box 2. BNF Code Structure.

Example presentation: Tradorec XL Tablets 300mg

Chapter	Section	Paragraph	Sub-paragraph	Chemical	Product	Presentation	Generic Equivalent
04	07	02	0	40	BI	AC	AM
Central Nervous System	Analgesics	Opioid Analgesics	Opioid Analgesics	Tramadol Hydrochloride	Tradorec	Tradorec XL_Tab 300mg	*

*Generic equivalent allows matching with the strength and formulation (presentation) of the generic product (which will always have product code 'AA').
For generic presentations the *product* name will match the *chemical* name (but sometimes with a different abbreviation, e.g. "Tramadol HCl").

In the PCA data, only the first seven characters of the BNF code for each drug are supplied, rather than the full BNF code. Therefore,the "drug name" is the only source of information on the formulation and dose ; however, from this the BNF code can usually be imputed, but this becomes increasingly difficult for older drugs no longer listed in the BNF. Although each drug's chemical name is also supplied, chemicals are not all unique (e.g. "Other Preparations"); names may change their spelling over time; and chemicals may move between Paragraphs, Sections and Chapters. Indeed, classifications at any level of the hierarchy can be subject to renaming, spelling change, subdivision, reorganisation and removal.

Data management, aggregation, and cleaning

All data were grouped by drug name, combining those differing only by standard quantity unit (SQU). Ultimately, following cleaning, data were grouped to product level. Medical devices/appliances and any other items Chapter numbers above 15 were excluded.

A key user-need was to explore prescribing trends for individual members of a class of drugs over time. This required all data to be normalised, with each individual drug consistently appearing in the correct location in the data schema; i.e. all individual presentations of a chemical all mapped under that chemical; and all chemicals mapped under the correct Sub-paragraph/Paragraph (often similar to drug class) of the BNF. To achieve this consistency, we aimed to map each drug to its current position in the latest BNF dictionary, up to the level of its 11-character "product" code, through an incremental process. This is summarised below and in Figure 1.

Lacking the full BNF code, we attempted to match each drug name to a current BNF presentation. Those without an exact match (e.g. formulation variants no longer available) could sometimes be matched to a similar BNF presentation name, e.g. by finding a similar formulation or using the "fuzzy" lookup add-on for Excel and validated manually [12]. Other drug names could only be matched up to current BNF codes by using their product or chemical names. Matching at each stage was improved by disregarding capitalisation, or spacing and spelling changes (e.g. Sulphur/Sulfur); these include changes identified within the data and those occurring when many old British spellings (the "British Approved Name") were replaced with international standard names (the "Recommended International Non-Proprietary Name") [13]. Remaining drug names in the most-prescribed Chapters (1-6 and 10) were matched to current drug names manually (for example, resolving non-matches due to rearrangement of word order); any others kept original chemical name, and a proxy product name was derived from the drug name field. Full methodology for this matching process is available in our technical documentation online [14] and in Supplementary Material.

We measured the extent of normalisation of drug names and classifications, and present summary statistics on these.

Normalisation for inflation and population

Prescribing costs were corrected for inflation using the UK's annual consumer price index (CPI) figures, normalised to 2016 [15]. Number of items prescribed and costs were divided by the population each year to calculate values per thousand population, based upon mid-

year population estimates for England only [16]. We also supply the original number of items and cost in our output.

Interactive Analysis Tool

Having generated a normalised dataset, and a method for updating it, we then set out to implement a free, interactive online data analysis tool where any user can visually explore time trends in prescribing. This was built using Tableau Public, a freely accessible interactive data presentation platform which permits rapid prototyping; however other front-ends onto the same underlying datasets could also be implemented using open source tools such as Shiny [17] or in Python libraries such as d3 [18] with more software engineer resource. Our user-needs for the tool were as follows: the ability to display trends in items and cost, normalised for total population change and inflation; and also to calculate the average cost per item and quantity per item for each product.

Having delivered the tool, we used it to generate trends data and graphs for a range of clinical areas where prescribing trends have been previously studied and published, to demonstrate the ability of our tool to replicate and extend these works.

Data and code

The full compiled and processed PCA datasets are available online via FigShare [19], SQL code is in Supplementary File, and the Trends tool (Tableau workbook) is available to use via https://OpenPrescribing.net/pca.

RESULTS

Data compilation and overall prescribing trends

All data were successfully imported. There were 169,100 lines of data in the compiled 1998-2016 dataset (Chapters 1-15) and 169,038 in the processed data, the reduction caused by aggregation of a small number of drugs available in multiple formulations despite having identical names. Total items (14.8 billion), cost (£136.9 billion) and distinct drug names (22,496) remained consistent before and after data processing (Table 1, S1). The inclusion of low volume prescribing in the published datasets from 2010 caused a substantial rise in the number of distinct drugs per year, but not items or cost (Table 1). As can be seen from Table 1, the inflation-corrected cost in 2016 UK sterling equivalent for all prescribing in NHS England primary care rose from £6.3bn in 1998 to £10.1bn in 2004, but then decreased to £8.3bn in 2016. Items per 1,000 population (correcting for crude population growth) has grown from 10,180 in 1998 to 19,196 in 2016, on average increasing by 3.6% per year.

Table 1. Summary of processed PCA data by year (drugs in Chapters 1-15 only). 'Drug Name' is the field describing the presentation of each drug, i.e. its formulation, dose and product name. Costs represent Net Ingredient Cost (NIC, see Box 1). 'Change' is the year-on-year change.

Year	Distinct count of Drug Name	Items	Items per 1000		Cost	Inflation-Corrected Cost	Inflation-Corrected Cost per 1000	
	n	n	n	change (%)	£	2016 £	2016 £	change (%)
1998	6,338	497.0M	10,180		£4,440M	£6,280M	£128,626	
1999	6,587	513.4M	10,471	+2.9%	£5,011M	£6,999M	£142,746	+11.0%
2000	6,613	535.1M	10,868	+3.8%	£5,284M	£7,318M	£148,648	+4.1%
2001	6,754	569.2M	11,510	+5.9%	£5,784M	£7,914M	£160,049	+7.7%
2002	6,834	598.6M	12,050	+4.7%	£6,487M	£8,768M	£176,491	+10.3%
2003	6,893	630.3M	12,625	+4.8%	£7,113M	£9,488M	£190,035	+7.7%
2004	6,912	666.0M	13,268	+5.1%	£7,645M	£10,063M	£200,482	+5.5%
2005	6,907	698.8M	13,808	+4.1%	£7,452M	£9,609M	£189,875	-5.3%
2006	6,810	728.4M	14,292	+3.5%	£7,660M	£9,655M	£189,436	-0.2%
2007	7,056	771.8M	15,022	+5.1%	£7,810M	£9,614M	£187,112	-1.2%
2008	7,202	816.7M	15,762	+4.9%	£7,716M	£9,174M	£177,047	-5.4%
2009	7,401	859.2M	16,461	+4.4%	£7,892M	£9,176M	£175,805	-0.7%
2010	11,703	898.4M	17,065	+3.7%	£8,162M	£9,193M	£174,636	-0.7%
2011	11,751	931.6M	17,541	+2.8%	£8,101M	£8,734M	£164,457	-5.8%
2012	12,207	968.9M	18,112	+3.3%	£7,802M	£8,176M	£152,836	-7.1%
2013	12,318	996.2M	18,494	+2.1%	£7,846M	£8,022M	£148,920	-2.6%
2014	12,576	1,027.0M	18,908	+2.2%	£8,022M	£8,078M	£148,718	-0.1%
2015	12,875	1,043.5M	19,046	+0.7%	£8,403M	£8,461M	£154,444	+3.9%
2016	13,285	1,060.9M	19,196	+0.8%	£8,284M	£8,284M	£149,892	-2.9%
Total	**22,496**	**14,810.9M**	**284,680**	**+3.6%**	**£136,914M**	**£163,006M**	**£3,160,254**	**1.0%**

Data Normalisation

Data were normalised using the methods described above. Of the distinct drug names in the data, 87.5% were matched exactly to a current BNF name, and a further 6.5% matched approximately (Table 2). Name changes are particularly prevalent in Chapter 3 (Respiratory), due mainly to the addition in 2004 of a space when a number of doses is given, as is common for inhalers, e.g. “Salbutamol_Inha 100mcg (200D)” became “Salbutamol_Inha 100mcg (200 D)”.

Table 2. Number and percentage of drug names subject to changes within 1998-2016 PCA data when compared to the current BNF, by Chapter. These include changes in word order, spacing, capitalisation, abbreviation, punctuation (e.g. “Califig_(California Syr Of Figs)”/“Califig_California Syr Of Figs”), spelling (e.g. “Sulphate”/“Sulfate”), brand name (e.g. “Laxoberal_Liq”/ “Dulcolax Pico_Liq”) and formulation (e.g. “Castor Oil_”/“Castor Oil_Liq”). The total count of drug names is reduced compared to Table 1 because the same drugs can appear over multiple years but only rarely in multiple Chapters.

Current Chapter code (name)	Name/spelling change		No change		No match		Grand Total
	n	%	n	%	n	%	n
1 (Gastro–intestinal system)		0.0%	1,041	99.1%	9	0.9%	**1,050**
2 (Cardiovascular system)	177	7.2%	2,278	92.5%	8	0.3%	**2,463**
3 (Respiratory system)	238	20.8%	893	78.0%	14	1.2%	**1,145**
4 (Central nervous system)	341	8.8%	3,535	90.7%	21	0.5%	**3,897**
5 (Infections)	119	9.1%	1,184	90.7%	3	0.2%	**1,306**
6 (Endocrine system)	203	13.2%	1,320	86.0%	11	0.7%	**1,534**
7 (Obstetrics, gynaecology and urinary–tract disorders)	4	0.7%	508	90.7%	48	8.6%	**560**
8 (Malignant disease and immunosuppression)	2	0.3%	538	91.8%	46	7.8%	**586**
9 (Nutrition and blood)	210	4.0%	4,281	81.3%	776	14.7%	**5,267**
10 (Musculoskeletal and joint diseases)	89	8.5%	956	90.8%	8	0.8%	**1,053**
11 (Eye)	18	3.0%	532	87.2%	60	9.8%	**610**
12 (Ear, nose and oropharynx)	16	3.4%	373	78.5%	86	18.1%	**475**
13 (Skin)	22	1.1%	1,772	89.5%	186	9.4%	**1,980**
14 (Immunological products and vaccines)	7	3.0%	198	85.0%	28	12.0%	**233**
15 (Anaesthesia)	15	5.2%	229	78.7%	47	16.2%	**291**
Grand Total	1,461	6.5%	19,638	87.5%	1,351	6.0%	**22,450**

Of the distinct drug names (23,275, taking into account some drugs having multiple BNF classifications), over 91% could be matched to a current product in the BNF, with no change in code (Table 3). Less than 5% could not be matched to a current product and/or chemical code, under 1% of items prescribed. These drugs were assigned proxy product names (derived from their drug name) so that all data could be presented visually, and for those not matched to a current chemical, the original chemical name was used (mostly "Other Preparations"). However, normalisation was focused on seven of the most prescribed Chapters with the greatest medical interest (1-6 and 10). The normalisation of drugs in other Chapters could therefore potentially be improved. In particular, Chapters 9 (Nutrition) and 13 (Skin) have substantial levels of prescribing, but are complex, containing many different drug names and non-drug products such as topical applications and dietary supplements. Other groups with a particular interest in nutrition or dermatology may wish to expand our work on manual matching: we would be happy to incorporate such amendments into our dataset. Many of the code changes and non-matches have diminished over time, as expected (Table S2).

Code changes and normalisation outputs are described in Table 3. Headers indicate the highest level in the BNF hierarchy at which drugs have been subject to code changes, e.g. "Section" indicates drug names which have not changed Chapter but have moved Section. "No product match" indicates drug names matched to a chemical (9-character BNF) but with no current matching product (11-character). "No chemical match" indicates drug names matched neither to a chemical nor product. The total count of drug names increases when separated by Chapter because four drug names currently exist in two different Chapters.

Table 3. Summary of drug code changes within the 1998-2016 prescribing datasets, also separated by (current) Chapter. Chapter names can be found in Table 2.

	BNF code change							
	Chapter	**Section**	**Paragraph**	**Sub-paragraph**	**No change**	**No product match**	**No chemical match**	**Grand Total**
Distinct count of Drug Name	94	52	560	203	21,258	815	293	**23,275**
% of Drugs	0.40%	0.22%	2.41%	0.87%	91.33%	3.50%	1.26%	**100%**
% of Items	0.04%	0.01%	0.84%	0.51%	97.67%	0.84%	0.10%	**100%**
Distinct count of Drug Name by Current Chapter								
Current Chapter number	**Chapter**	**Section**	**Paragraph**	**Sub-paragraph**	**No change**	**No product match**	**No chemical match**	**Grand Total**
1		4	109		1,146	6		**1,265**
2	5	3	4	19	2,446	7		**2,484**

3	1	6	19		1,132	10		**1,168**
4	42	13	173		3,831	9		**4,068**
5	1		161	116	1,241	2		**1,521**
6	4	1	2	2	1,521	10		**1,540**
7	5		12		528	28		**573**
8	1		4		579	6	1	**591**
9	1	7	6	54	4,510	507	245	**5,330**
10	27				1,041	9		**1,077**
11	1	5	5		571	28	6	**616**
12					420	44	11	**475**
13	5	13	28	12	1,847	95	30	**2,030**
14			37		195	17		**249**
15	1				254	37		**292**
Grand Total	**94**	**52**	**560**	**203**	**21,262**	**815**	**293**	**23,279**

Interactive Data Analysis Tool

We created a tool which allows anyone to explore the prescribing data, available directly at https://openprescribing.net/pca. Users can search by chemical, Paragraph, Section or Chapter to view time trends in items and costs on stacked charts, where both the overall trends and the relative contribution from each product/chemical can be seen. The cost per item and quantity per item for each product are also shown, which can assist in interpretation of trends in some cases. However, these calculations carry a “use with caution” note, as items may represent different pack sizes, and quantities cannot be reliably summed across preparations because of different strengths and formulations. The page features an accompanying video walk-through demonstrating the tool.

The tool can be used to facilitate novel research into time trends, and factors associated with changes in practice such as publication of guidelines or evidence landmarks, or changes in price. It can replicate and extend the main findings of previous papers which researched trends for different clinical areas using PCA data. For example, the antipsychotic drug switches which occurred in England following a licence restriction [20] can be replicated in the tool and the trends extended to the latest data (Figure 2a). This also shows the dramatic reductions in cost that followed the expiry of patents for risperidone and olanzapine. We also replicate antidepressant prescribing trends, previously reported up to 2010 [5], and show that how the overall use of these antidepressants has continued to rise, in particular sertraline (Figure 2b). We also replicate findings on the rise of thyroid hormones [3] and testosterone [4], where we show that prescribing of these drugs continued to rise, with a disproportionate increase in cost (Figure S1a-b). We are using this dataset and tool in our

academic papers on trends and variation in NHS prescribing; we encourage others to use our dataset and tool in their own work.

The tool can also be used to complement studies performed in more detailed prescribing data such as the Clinical Practice Research Datalink (CPRD), by giving the full national picture, and giving more longitudinal data that updates with new data releases. For example, several previous publications have reported on patterns of prescribing of smoking cessation medication in The Health Improvement Network (THIN) database [21–23]. This included reporting of a possible decline in prescribing despite increased incentives for GPs introduced in 2012. We can confirm this decline and show that it continued beyond 2013 (Figure 2c). We also show that the slow decline in quinine usage following safety alerts in 2010 [24] has continued at a similar pace (Figure S1c). CPRD data contain individual patient records and can therefore be used to assess detailed questions about treatments in specific cohorts of patients. However many labour-intensive CPRD analyses have been conducted to interrogate simple broad prescribing trends which could more straightforwardly be conducted using aggregated and normalised national data, with greater coverage of years and total population. Furthermore, for analyses interrogating national trends and responses to guidelines, in many cases a prescribing change which can only be detected in individual patients' records, and cannot be detected in national data, may not be relevant in terms of population health or the health service.

Additional tabs in the tool allow discovery of higher-level trends, including Chapter and Section trends, Sections ranked by items/cost for any selected year, calculation of the change in items/cost for each Section between any selected year to the latest year, and the top 20 Paragraphs by items and cost. The Chapter trends page, for example, shows that much of the decline in prescribing costs since the peak in 2004 (Table 1) is attributable to a drop in the cost of cardiovascular drugs (Figure 3a), and the Section trends page further shows that lipid-regulating drugs (Section 2.12) and Drugs for Hypertension and Heart Failure (Section 2.5) experienced the largest cost reductions at that time (Figure 3b).

DISCUSSION

Summary

It was possible to aggregate all PCA data from 1998-2016 and normalise for most changes in drug names and classifications. Only 87.5% of drug names matched exactly to a current BNF name and 8.7% had undergone some change in classification; however all drugs in core clinical Chapters were reconciled to their current location in the data schema. We

generated an interactive online service where any user can explore time trends in prescribing broken down by product, chemical, Paragraph, Section and Chapter; this openly accessible interactive data analysis tool provides overviews and insights comparable to previous labour-intensive bespoke data analysis research projects.

Strengths and weaknesses of this study

Our tool covers the data for the whole of England's community dispensing, not a sample. We are surprised to note that this is the first project aiming to aggregate long-term trends across the entire prescribing dataset, and provide an openly accessible tool for wider use. Many drugs changed name and/or classification over time, but valid chemicals were successfully assigned to all items in Chapters 1-6 and 10, and product names were derived for every drug, allowing maximum consistency in trends analysis. The tool is limited to product-level data due to the wide number of different presentations available.

We used items to measure prescribing volume. Quantity is generally more complex for making comparisons as there is wide variation caused by the units, which may be the number of pills or millilitres, units (such as inhalers containing multiple doses) or other unit measure. Converting quantities to approximate daily dose sizes (such as Defined Daily Doses, DDDs) is possible, but the conversion tables available are not sufficiently comprehensive to allow this across the entire dataset, and this would be even more difficult for discontinued drugs. We therefore rejected this option in favour of being able to publish a complete dataset. However, users wishing to analyse data by daily doses can download our BNF-normalised dataset in order to apply these calculations. Using items also has limitations, as it does not take into account number of packs prescribed per prescription, pack size or dosage. We are launching this tool publicly and will monitor user volume and user-feedback: if appropriate we will improve the tool by replicating and expanding it using bespoke software as per our other data analysis tools on OpenPrescribing.net for exploring variation in prescribing at CCG and individual practice level. We will update the tool annually, dependent upon continuing funds for the OpenPrescribing project.

Findings in context of other research

Long-term trends in prescribing have previously been reported on a wide variety of clinical areas, using PCA data as well as other sources [3–5,20,22,23]. These are static, not updated, and rapidly out of date. Although using CPRD allows a more detailed analysis and investigation of patient factors associated with prescribing, it takes a great deal of preparation and time to complete. Our tool can replicate some trends found in CPRD, and so may provide a useful tool for preliminary investigation of trends. It can also help to confirm

whether findings from regional datasets of rich individual patient data (IPD) from electronic health records sources are representative of the national picture, while avoiding repeated work and replication in new IPD datasets. In our related publications on variation and trends in specific disease areas we report comparisons between trends in PCA data, and trends from other more labour-intensive sources such as CPRD, in more detail.

The UK government produces a 10-year trends document following the annual PCA data release, containing an overall summary of high-level trends and a brief breakdown of six interesting topics with the greatest level or change in prescribed items and cost [25]. However, the reported topics are few in number, chosen by NHS Digital, restricted to ten years of data, do not correct for inflation or population growth, are not easily discoverable by subject specialists, and readers are not able to interrogate their own topics of interest in detail. From 2016, the compiled datasets were also made available so users may conduct their own exploration of the data, but, without drug names or categories being normalised, this is little better than the raw data, which we have processed into a normalised longitudinal dataset.

Policy implications and future research

Published papers can provide a useful and detailed insight into prescribing trends [3–5], but give a single snapshot which may quickly become out of date. Our tool facilitates ongoing monitoring by researchers and policymakers to assess prescribing changes in any area of concern or clinical interest they have identified; and permits interactive exploration of detailed issues in the data, such as individual presentations of chemicals, by any interested user. As part of our OpenPrescribing work we are using prescribing data to investigate adherence to guidelines and changes in practice in various clinical areas, to detect anomalous changes in individual practices relative to national trends in order to send practices alerts, and to identify cost saving opportunities. We have produced various manuscripts using the longitudinal data presented here as part of a range of data sources to describe variation in prescribing in primary care. We are happy to collaborate with other teams of clinicians and academics; we also release our underlying dataset and code as open data for re-use with citation.

Conclusions

Long-term trends in prescribing are interesting for a number of applications. While previous work on prescribing data has focused on static, manual analysis of a small number of drugs, modern data science approaches make it possible to create interactive services that allow clinicians, healthcare commissioners, policy makers, academics and any other interested

party to interrogate and monitor prescribing trends for any combination of chemicals, to identify anomalies or signals of concern, and predict spending. We have delivered this using a combination of open data and freely accessible online tools.

ACKNOWLEDGEMENTS

We are grateful to Seb Bacon for maintaining databases and general assistance, and to Richard Croker for pharmaceutical advice and fuzzy matching in Excel.

FUNDING

OpenPrescribing is funded by: NIHR Biomedical Research Centre Oxford, Health Foundation (Ref 7599), NIHR SPCR (Ref 327). No specific funding was sought for this project. The authors' funders had no involvement in the study design or the decision to submit.

CONFLICT OF INTEREST

All authors have completed the Unified Competing Interest form (available on request from the corresponding author) and declare: BG has received research funding from the Laura and John Arnold Foundation, the Wellcome Trust, the NHS NIHR School of Primary Care, the Health Foundation, NHS England, NIHR Biomedical Research Centre Oxford, and the WHO; he also receives personal income from speaking and writing for lay audiences on the misuse of science. HC is employed on BG's OpenPrescribing grants.

CONTRIBUTORSHIP STATEMENT

BG conceived and supervised the project, HC designed the methods, conducted the analysis, interpreted the findings, extracted and processed the data in BigQuery, Excel and Tableau with input from BG. HC and BG wrote the paper. All authors contributed to and approved the final manuscript. BG is guarantor.

DATA SHARING STATEMENT

The full compiled and processed PCA datasets are available online via Figshare [DOI 10.6084/m9.figshare.5447194.v1], SQL code is in Supplementary File, and the Trends tool (Tableau workbook) is available to use via https://OpenPrescribing.net/pca.

REFERENCE LIST

1 NHS-Digital. National Statistics Prescription Cost Analysis, England - 2016. 2016.http://www.content.digital.nhs.uk/searchcatalogue (accessed 28 Sep 2017).

2 NHS-Choices. About the National Health Service (NHS) in England - NHS Choices. 2016.http://www.nhs.uk (accessed 28 Sep 2017).

3 Mitchell AL, Hickey B, Hickey JL, *et al.* Trends in thyroid hormone prescribing and consumption in the UK. *BMC Public Health* 2009;**9**:132.

4 Gan EH, Pattman S, H S Pearce S, *et al.* A UK epidemic of testosterone prescribing, 2001-2010. *Clin Endocrinol* 2013;**79**:564–70.

5 Ilyas S, Moncrieff J. Trends in prescriptions and costs of drugs for mental disorders in England, 1998–2010. *Br J Psychiatry* 2012;**200**:393–8.

6 Dietrich ES. Effects of the National Institute for Health and Clinical Excellence's technology appraisals on prescribing and net ingredient costs of drugs in the National Health Service in England. *Int J Technol Assess Health Care* 2009;**25**:262–71.

7 Connor AJ, Fraser SG. Glaucoma prescribing trends in England 2000 to 2012. *Eye* 2014;**28**:863–9.

8 Fitzpatrick RW, Pate RG. Assessing the impact of NICE guidance on the prescribing of hormonal treatments of breast cancer in England. *J Eval Clin Pract* 2015;**21**:759–61.

9 NHS-Digital. Prescribing Data. 2017.https://digital.nhs.uk/article/4214/Prescribing (accessed 21 Sep 2017).

10 NHS-Digital. Prescription Cost Analysis - England, 2009 [NS]. 2012.https://digital.nhs.uk/catalogue/PUB01414 (accessed 11 Sep 2017).

11 NHS-Digital. Prescription Cost Analysis Glossary 2015. pres-cost-anal-eng-2015-apx.pdf. 2016.http://content.digital.nhs.uk/catalogue/PUB20200/pres-cost-anal-eng-2015-apx.pdf (accessed 28 Sep 2017).

12 Curtis HJ. Prescription Cost Analysis 1998-2016 - drugs matched to BNF via fuzzy lookup. *figshare* Published Online First: 28 September 2017. doi:10.6084/m9.figshare.5450323.v1

13 Tidy C. Name Changes of Medicines. Patient Platform Limited. 2015.https://patient.info/health/name-changes-of-medicines (accessed 20 Sep 2017).

14 Curtis HJ. Prescription Cost Analysis 1998-2016 data processing and normalisation. 2017.https://gist.github.com/HelenCEBM/192307b3c671a391f5ad6b44a3676880 (accessed 28 Sep 2017).

15 Office for National Statistics. Population Estimates. ONS Statistical bulletin. 2017.https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2016 (accessed 11 Sep 2017).

16 Office for National Statistics. CPI All Items Index. ONS Time Series Data. 2017.https://www.ons.gov.uk/economy/inflationandpriceindices/timeseries/d7bt/mm23 (accessed 20 Sep 2017).

17 RStudio-Inc. Shiny. 2017.https://shiny.rstudio.com/ (accessed 27 Sep 2017).

18 Bostock M. D3.js - Data-Driven Documents. 2017.https://d3js.org/ (accessed 27 Sep 2017).

19 Curtis HJ, Goldacre B. NHS prescription cost analysis data 1998-2016. *figshare* Published Online First: 28 September 2017. doi:10.6084/m9.figshare.5447194.v1

20 Bateman DN, Good AM, Afshari R, *et al.* Effects of licence change on prescribing and poisons enquiries for antipsychotic agents in England and Scotland. *Br J Clin Pharmacol* 2003;**55**:596–603.

21 Szatkowski L, Coleman T, McNeill A, *et al.* The impact of the introduction of smoke-free legislation on prescribing of stop-smoking medications in England. *Addiction* 2011;**106**:1827–34.

22 Langley TE, Huang Y, McNeill A, *et al.* Prescribing of smoking cessation medication in England since the introduction of varenicline. *Addiction* 2011;**106**:1319–24.

23 Szatkowski L, Aveyard P. Provision of smoking cessation support in UK primary care: impact of the 2012 QOF revision. *Br J Gen Pract* 2016;**66**:e10–5.

24 Acheampong P, Cooper G, Khazaeli B, *et al.* Effects of MHRA drug safety advice on time trends in prescribing volume and indices of clinical toxicity for quinine. *Br J Clin Pharmacol* 2013;**76**:973–9.

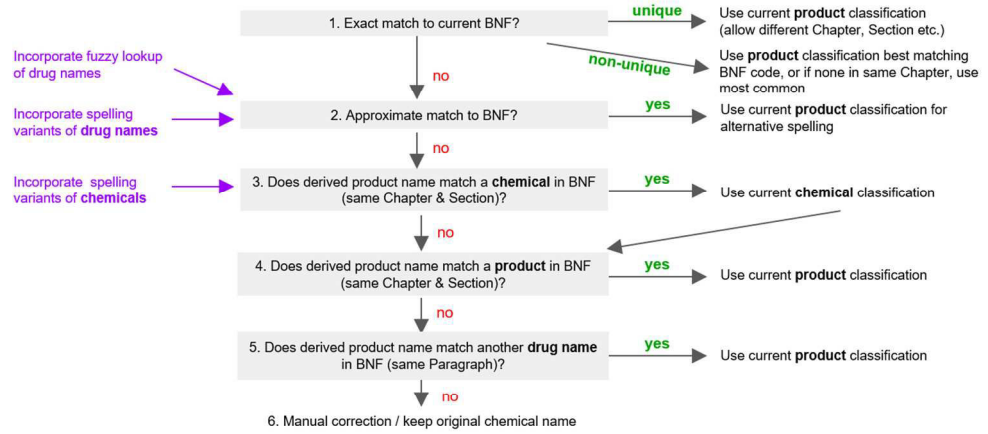
25 NHS-Digital. Prescriptions Dispensed in the Community, England 2006 to 2016. *Prescribing and Medicines Team, NHS Digital, UK* Published Online First: 29 June 2017.https://digital.nhs.uk/catalogue/PUB30014

LEGENDS TO FIGURES

Figure 1. *BNF code normalisation process flow chart describing how drug names were matched to the current BNF. Examples of matches at each numerated stage are given in table (bottom), with code/name changes in red. The first example demonstrates a drug which was matched to the current BNF through an exact match by name, but had moved from Chapter 4 to Chapter 2. For other types of matching, care was taken to avoid mistakenly matching to similar (but distinctly different) drugs across Chapters/Sections.*

Figure 2. *Screenshots from Trends tool, showing items per 1,000 population and inflation-corrected costs per 1,000 population for selected drugs. Product names are in parentheses. (a) Prescribing trends for four selected antipsychotic chemicals following the safety alert on thioridazine [20]. Full dashboard available at https://public.tableau.com/shared/XX7DTWSG2?:display_count=yes. (b) Prescribing trends for selected antidepressant chemicals [5]. Full dashboard available at https://public.tableau.com/shared/72SJGGP89?:display_count=yes. (c) Prescribing trends for all chemicals within the Paragraph of Nicotine Dependence (smoking cessation medications). Full dashboard available at https://public.tableau.com/shared/6BW9J5RJB?:display_count=yes.*

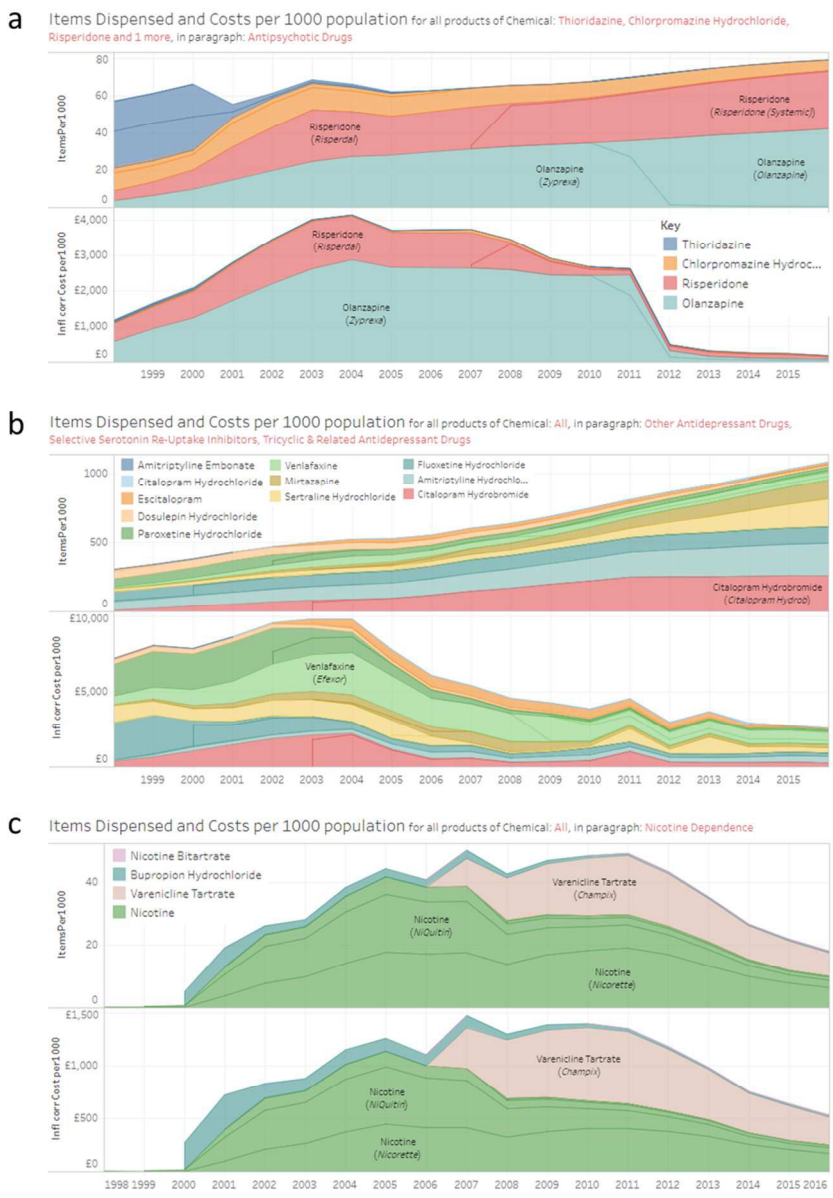
Figure 3. *Screenshots from Trends tool, showing inflation-corrected costs per 1,000 population (a) by Chapter, and (b) by Section for Chapter 2 (Cardiovascular System).*



Match Type	Drug Name Original	Chemical Original	BNF Code Original	Drug Name Current	Chemical Current	Code Derived	Product Current
1	Aspirin_Pdr Sach 37.5mg	Aspirin	0407010	Aspirin_Pdr Sach 37.5mg	Aspirin	0209000A0AA	Aspirin (Antiplatelet)
2	Frusemide_Tab 20mg	Furosemide	0202020	Furosemide_Tab 20mg	Furosemide	0202020L0AA	Furosemide
2	Amlodipine Besyl_Liq Spec 10mg/5ml	Amlodipine Besylate	0206020	Amlodipine_Oral Soln 5mg/5ml	Amlodipine	0206020A0AA	Amlodipine
3	Felendil XL_Tab 10mg P/R	Felodipine	0206020	(none)	Felodipine	0206020F0	Felendil XL
4	Cozaar-Comp_Tab	Losartan Potassium With Diuretic	0205052	(none)	Losartan Potassium With Diuretic	0205052P0BB	Cozaar-Comp
5	Atorvastatin_Liq Spec 20mg/5ml	Atorvastatin	0212000	Atorvastatin_Oral Soln 20mg/5ml	Atorvastatin	0212000B0AA	Atorvastatin
6	Terbut Sulph_Syr 1.5mg/5ml S/F	Terbutaline Sulphate	0301011	(none)	Terbutaline Sulfate	0301011V0AA	Terbut Sulf

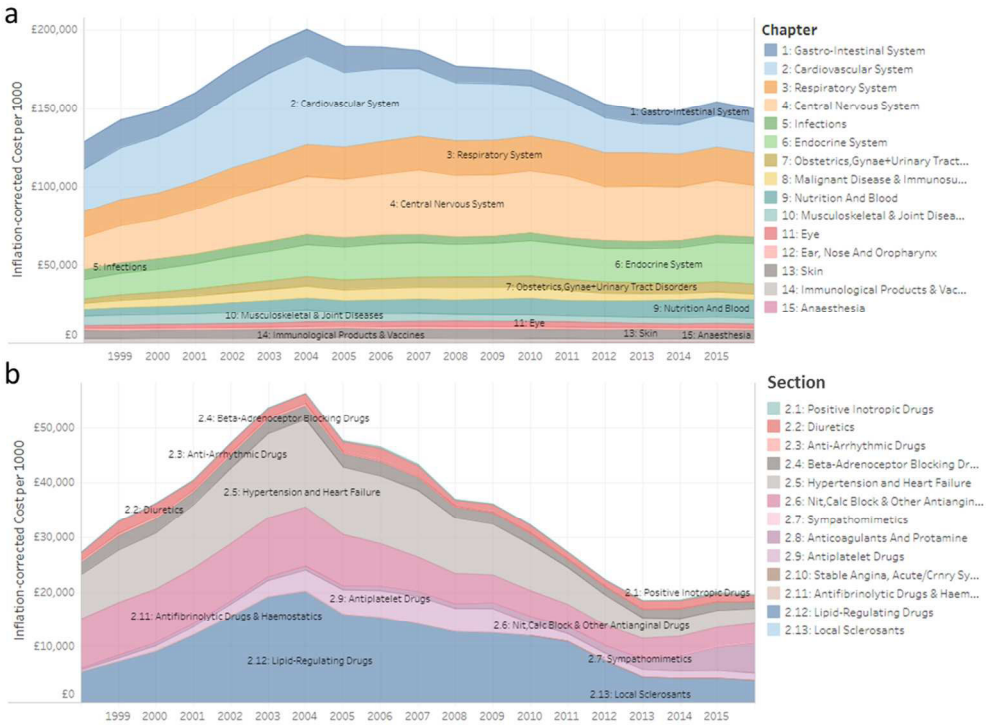
BNF code normalisation process flow chart describing how drug names were matched to the current BNF. Examples of matches at each numerated stage are given in table (bottom), with code/name changes in red. The first example demonstrates a drug which was matched to the current BNF through an exact match by name, but had moved from Chapter 4 to Chapter 2. For other types of matching, care was taken to avoid mistakenly matching to similar (but distinctly different) drugs across Chapters/Sections.

116x70mm (300 x 300 DPI)



Screenshots from Trends tool, showing items per 1,000 population and inflation-corrected costs per 1,000 population for selected drugs. Product names are in parentheses. (a) Prescribing trends for four selected antipsychotic chemicals following the safety alert on thioridazine [20]. Full dashboard available at https://public.tableau.com/shared/XX7DTWSG2?:display_count=yes. (b) Prescribing trends for selected antidepressant chemicals [5]. Full dashboard available at https://public.tableau.com/shared/72SJGGP89?:display_count=yes. (c) Prescribing trends for all chemicals within the Paragraph of Nicotine Dependence (smoking cessation medications). Full dashboard available at https://public.tableau.com/shared/6BW9J5RJB?:display_count=yes.

127x179mm (300 x 300 DPI)



Screenshots from Trends tool, showing inflation-corrected costs per 1,000 population (a) by Chapter, and (b) by Section for Chapter 2 (Cardiovascular System).

127x91mm (300 x 300 DPI)

SUPPLEMENTARY FILE

OpenPrescribing: Normalised Data and Software Tool to Research Trends in English NHS Primary Care Prescribing 1998-2016

Contents:

Table S1. Summary of original (unprocessed) PCA data by year (drugs in Chapters 1-15 only).

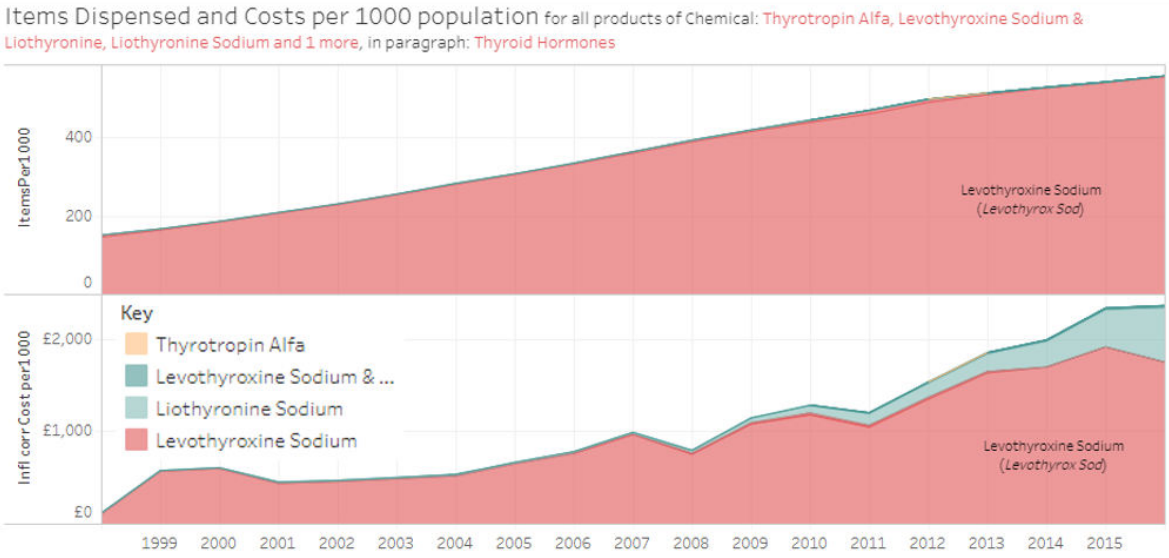
Year	Distinct count of Drug Name	Items	Cost
1998	6,338	497M	£4,440M
1999	6,587	513M	£5,011M
2000	6,613	535M	£5,284M
2001	6,754	569M	£5,784M
2002	6,834	599M	£6,487M
2003	6,893	630M	£7,113M
2004	6,912	666M	£7,645M
2005	6,907	699M	£7,452M
2006	6,810	728M	£7,660M
2007	7,056	772M	£7,810M
2008	7,202	817M	£7,716M
2009	7,401	859M	£7,892M
2010	11,703	898M	£8,162M
2011	11,751	932M	£8,101M
2012	12,207	969M	£7,802M
2013	12,318	996M	£7,846M
2014	12,576	1,027M	£8,022M
2015	12,875	1,043M	£8,403M
2016	13,285	1,061M	£8,284M
Total	***27,473***	**14,811M**	**£136,914M**

Table S2. Summary of drug code changes within the 1998-2016 prescribing datasets, by year.

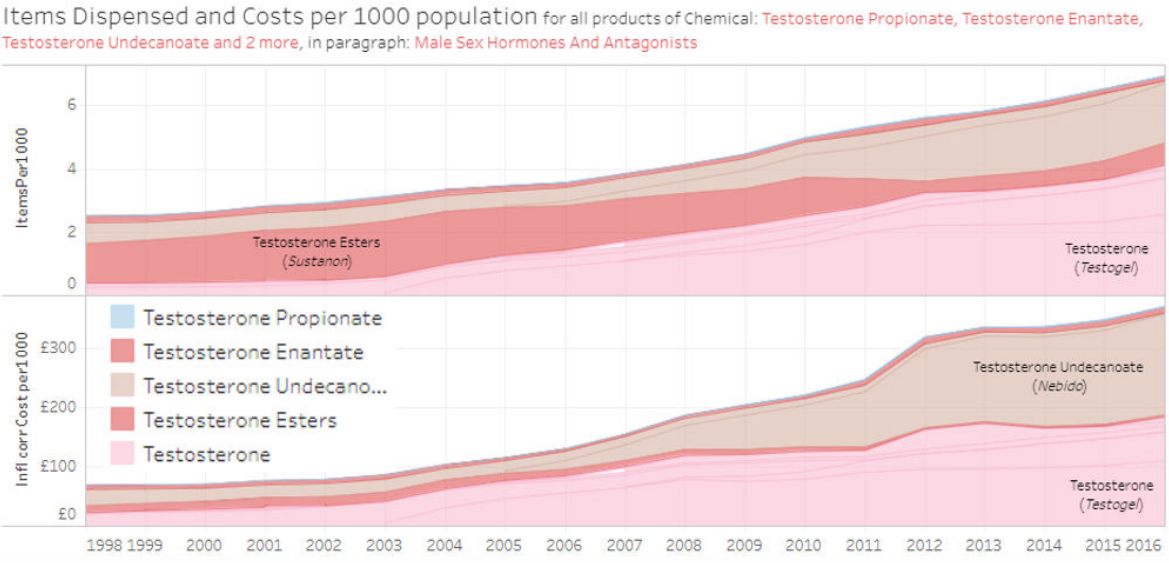
	Code change							
year	Chapter	Section	Paragraph	Sub-Paragraph	no change	no product match	no chemical match	Grand Total
1998	21	9	219	140	5,609	230	110	**6,338**
1999	21	7	188	88	5,917	237	129	**6,587**
2000	22	8	189	86	5,928	260	120	**6,613**
2001	26	9	178	85	6,069	278	111	**6,756**
2002	31	10	182	77	6,173	274	87	**6,834**
2003	34	9	191	79	6,256	241	83	**6,893**
2004	36	16	191	77	6,265	240	87	**6,912**
2005	26	14	157	85	6,315	245	65	**6,907**

2006	27	17	158	72	6,245	232	59	**6,810**
2007	35	12	168	74	6,483	237	47	**7,056**
2008	34	12	169	74	6,648	231	34	**7,202**
2009	3	4	180	76	6,919	219		**7,401**
2010	12	8	207	14	11,182	279	1	**11,703**
2011	15	4	130	14	11,331	257		**11,751**
2012	8	4	4	16	11,946	229		**12,207**
2013	2	4	8	1	12,139	163	1	**12,318**
2014		6	15	1	12,415	139		**12,576**
2015		6	6		12,770	93		**12,875**
2016		6	7		13,241	31		**13,285**

a


Items Dispensed and Costs per 1000 population for all products of Chemical: Thyrotropin Alfa, Levothyroxine Sodium & Liothyronine, Liothyronine Sodium and 1 more, in paragraph: Thyroid Hormones
ItemsPer1000
400
200
0
Levothyroxine Sodium
(Levothyrox Sod)
Infl corr Cost per1000
£2,000
£1,000
£0
Key
Thyrotropin Alfa
Levothyroxine Sodium & ...
Liothyronine Sodium
Levothyroxine Sodium
Levothyroxine Sodium
(Levothyrox Sod)
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

b


Items Dispensed and Costs per 1000 population for all products of Chemical: Testosterone Propionate, Testosterone Enantate, Testosterone Undecanoate and 2 more, in paragraph: Male Sex Hormones And Antagonists
ItemsPer1000
6
4
2
0
Testosterone Esters
(Sustanon)
Testosterone
(Testogel)
Infl corr Cost per1000
£300
£200
£100
£0
Testosterone Propionate
Testosterone Enantate
Testosterone Undecano...
Testosterone Esters
Testosterone
Testosterone Undecanoate
(Nebido)
Testosterone
(Testogel)
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

c

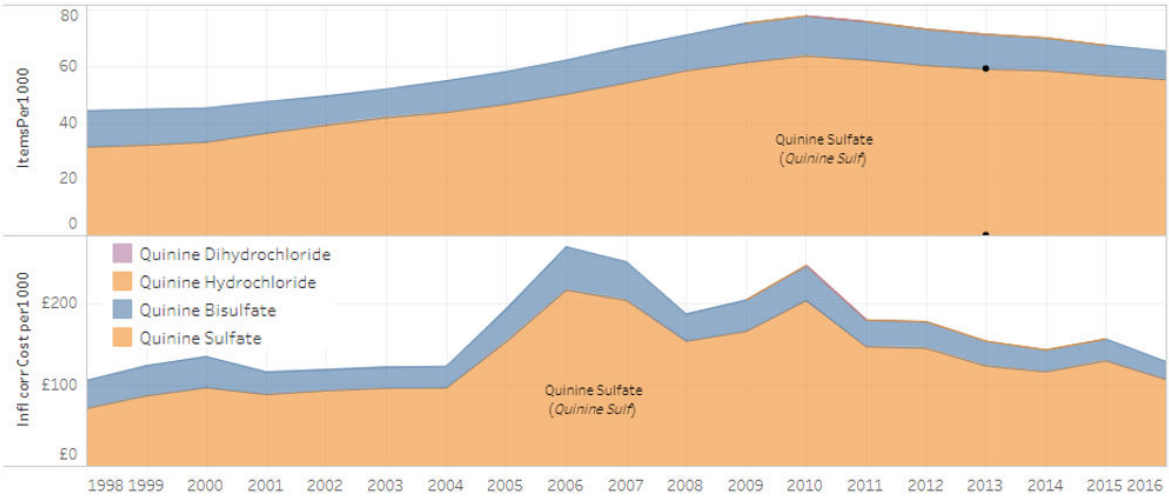

Items Dispensed and Costs per 1000 population for all products of Chemical: Quinine Dihydrochloride, Quinine Hydrochloride, Quinine Bisulfate and 1 more, in paragraph: Antimalarials
ItemsPer1000
80
60
40
20
0
Quinine Sulfate
(Quinine Sulf)
Infl corr Cost per1000
£200
£100
£0
Quinine Dihydrochloride
Quinine Hydrochloride
Quinine Bisulfate
Quinine Sulfate
Quinine Sulfate
(Quinine Sulf)
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

Figure S1. Screenshots from Trends tool, showing items per 1,000 population and inflation-corrected costs per 1,000 population for selected drugs. Product names are in parentheses. (a) Prescribing trends for all chemicals within the Paragraph of Thyroid Hormones. Full dashboard available at https://public.tableau.com/shared/GPW28PWJY?:display_count=yes. (b) Prescribing trends for all testosterone chemicals within the Paragraph of Male Sex Hormones. Full dashboard available at https://public.tableau.com/shared/YQ3ZFB3HY?:display_count=yes. c) Prescribing trends for all chemical forms of quinine (all of which are in the Antimalarials Paragraph). Full dashboard available at https://public.tableau.com/shared/85KJ2ZFN4?:display_count=yes.

Appendix - SQL Code for Processing and Normalisation of PCA data

A - Lookup Tables

A1. The special_cases lookup table

This is a workaround to assign a 'most likely' classification to the few problematic drug names which exist multiple times in BNF.

Lookup table is created by running the following script:

```
WITH temp as (
SELECT SUBSTR(SECTION_CODE,1,2) as chapter, section_code, presentation,
COUNT(DISTINCT product_code) as num
FROM ( SELECT DISTINCT section_code, section, para, subpara, chemical, product,
product_code, presentation FROM ebmdatalab.hscic.bnf )
GROUP BY chapter, section_code, presentation
HAVING num >1   --where name maps to more than one bnf code
ORDER BY chapter, num DESC)

SELECT
section_code, section, para, subpara, chemical, product, product_code,
presentation  -- this level is to filter to the top-prescribed code for each
drug name (according to latest detailed monthly data) (or, if none were
prescribed, then the first product name alphabetically
FROM (
  SELECT  -- this level joins all possible product codes to aggregated
prescribing data (2011-16) and ranks by items prescribed.
    a.*, b.items AS items_2011_2016,
    row_number() OVER (PARTITION BY a.presentation ORDER BY b.items DESC) AS
ranking -- We can use this to select the top/most likely drug code
  FROM (  -- this level is to look up all possible product codes for each drug
name in current BNF
      SELECT
```

```
      DISTINCT -- here we just want to go to product level rather than
individual presentations
      presentation,
      chapter, chapter_code,
      section, section_code,
      para, para_code,
      subpara, subpara_code,
      chemical, chemical_code,
      product, product_code
      from ebmdatalab.hscic.bnf where presentation in (select presentation from
temp where chapter < '18')

      ) a  -- now join to aggregated dataset grouped up to product level:
      LEFT JOIN ( SELECT substr(bnf_code,1,11) AS product_code, sum(items) as
items from ebmdatalab.aggregated_data.all_prescribed_BNFs_UpToSept2016 GROUP BY
product_code ) b
            ON a.product_code = substr(b.product_code,1,11)
      )
WHERE ranking = 1
ORDER BY chapter_code, presentation, product_code
```

A2. The lookup table of alternative drug spellings found within the data is created using the script below and saving as **ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_HC**

```
-- find drug name changes in PCA data to 2016
-- save results as ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_HC

WITH
a AS  (
  SELECT
      IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4)) AS
section_code, -- extra clause added to deal with those with extra spaces
2017-08-01

SUBSTR(drug_name,1,IF(STRPOS(drug_name,'_')>0,STRPOS(drug_name,'_')-1,length(dr
ug_name))) AS drug_name_part,  --take first part of drug name, up to underscore
(if there is one)
      MIN(year) AS min_year,  --this will help us to see which are the older vs
newer spellings used
      MAX(Year) AS max_year,
      SUM(items) AS Items
  FROM ebmdatalab.hscic.prescribing_pca_1998_2016_full
  WHERE  IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,2),SUBSTR(bnf_7_char,1,2))
< '18'
  GROUP BY
```

```
  section_code,
  drug_name_part
  ),

 b AS
(SELECT DISTINCT
section_code,
drug_name_part,
REPLACE(REPLACE(drug_name_part, 'i', '__'),'y','__') AS IY,
REPLACE(REPLACE(drug_name_part, 's', '__'),'z','__') AS SZ,
REPLACE(REPLACE(drug_name_part, 'ph', '__'),'f','__') AS PHF,
CONCAT(drug_name_part,'e') AS E,   -- add and E on to the end (note this only
works for the LAST word)
REPLACE(drug_name_part, ' ', ' e') AS E_mid -- add an E on to the end of all
words occurring before a space
FROM a)


SELECT
a.section_code,
a.drug_name_part,
CAST(a.min_year AS STRING) AS start_date,
CAST(a.max_year AS STRING) AS end_date,
a.items,
b.drug_name_part AS alternative,
CASE WHEN REPLACE(REPLACE(a.drug_name_part, 'i', '__'),'y','__') = b.IY THEN
'i_y'
      WHEN  REPLACE(REPLACE(a.drug_name_part, 's', '__'),'z','__') = b.SZ THEN
's_z'
      WHEN  REPLACE(REPLACE(a.drug_name_part, 'ph', '__'),'f','__') = b.PHF
THEN 'ph_f'
      WHEN  a.drug_name_part = b.E OR CONCAT(a.drug_name_part,'e') =
b.drug_name_part THEN 'e_end'
      WHEN  a.drug_name_part = b.E_mid OR REPLACE(a.drug_name_part, ' ', 'e ')
=b.drug_name_part THEN 'e_end'
      END AS type
FROM a
INNER JOIN b
  ON (REPLACE(REPLACE(a.drug_name_part, 'i', '__'),'y','__') = b.IY
      OR REPLACE(REPLACE(a.drug_name_part, 's', '__'),'z','__') = b.SZ
      OR REPLACE(REPLACE(a.drug_name_part, 'ph', '__'),'f','__') = b.PHF
      OR a.drug_name_part = b.E   --note this will only show this match once,
so we put in the other way around also
      OR a.drug_name_part = b.E_mid
      OR CONCAT(a.drug_name_part,'e') = b.drug_name_part
      OR REPLACE(a.drug_name_part, ' ', 'e ') =b.drug_name_part)
  AND a.drug_name_part != b.drug_name_part
  AND a.section_code = b.section_code
```

```
ORDER BY items desc
```

A3. The lookup table of Chemical name changes is created using the script below and saving as **pca_chemical_old_to_new_lookup**

```
-- PCA data - finding up to date chemical to combine with dataset
--save results as ebmdatalab.hscic.pca_chemical_old_to_new_lookup
WITH A as (
  SELECT
    IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4)) AS
Section,
    drug_name,
    count(distinct chemical) AS chems,
    max(year) AS Max_year_overall
    FROM
    ebmdatalab.hscic.prescribing_pca_1998_2016_full
    where
IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,2),SUBSTR(bnf_7_char,1,2)) <'18'
    GROUP BY
    Section,
    drug_name
    HAVING chems >1
    ),

B AS (
SELECT IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4))
AS section, drug_name, chemical,
    min(year) AS Min_year,
    max(year) AS Max_year
    FROM ebmdatalab.hscic.prescribing_pca_1998_2016_full
GROUP BY Section, drug_name, chemical
    ),

C AS (
SELECT DISTINCT
    A.Section,
    A.drug_name,
    B.chemical,
    B.min_year,
    B.max_year,
    IF(max_year = Max_year_overall,1,0) AS latest
    FROM A LEFT JOIN B ON A.drug_name = B.drug_name AND A.Section = B.Section
ORDER BY drug_name, chemical
)

SELECT old.section, old.drug_name, old.chemical AS old_chemical_name,
    nw.chemical AS new_chemical_name, nw.min_year AS Since
    FROM c old
```

```
     LEFT JOIN c nw ON old.drug_name = nw.drug_name AND old.chemical !=
nw.chemical and nw.latest = 1
    WHERE old.latest = 0
ORDER BY old.section, old.drug_name
```

A4. Known drug name changes

As reported online by patient.info
`ebmdatalab.hscic.drug_name_changes_2013`

A5. Fuzzy lookup for drugs not matching to BNF

List of drugs not matching BNF, identified through earlier iterations of the code.
These 1,084 drugs were matched to similar BNF names via fuzzy lookup in Excel and manually checked by a pharmacist.

List available at:
https://docs.google.com/spreadsheets/d/1UweKIZOLrKEzCtLULk5R5kJ4UyFIttvEouQ7RFGYrEA/edit#gid=594622641
and stored as `ebmdatalab.hscic.pca_bnf_name_to_code_fuzzy_lookup`

B - Data Extraction And Normalisation

B1. The latest chemical name for each drug is appended into the full dataset, to create prescribing_pca_1998_2016_full_v2

This does not take into account spelling changes but those will be handled later

```
-- save results as ebmdatalab.hscic.prescribing_pca_1998_2016_full_v2

SELECT a.*, COALESCE(c2.new_chemical_name, a.chemical) AS new_chemical_name
FROM
ebmdatalab.hscic.prescribing_pca_1998_2016_full a

LEFT JOIN ebmdatalab.hscic.pca_chemical_old_to_new_lookup c2
   ON a.drug_name = c2.drug_name
   AND a.chemical = c2.old_chemical_name
   AND IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4)) =
c2.section
```

C2. Run final data extraction parts 1 and 2 (scripts copied and updated from Issues #6 and #7)

B2a. Part 1

-- Final PCA data extraction part 1 (2016)

```
-- save results as ebmdatalab.tmp_eu.trends_from_pca

WITH
temp AS
(SELECT DISTINCT X.section_code, X.drug_name_part AS old_name, X.alternative
FROM
ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_HC X
INNER JOIN ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_2016_HC Y ON
X.alternative = Y.drug_name_part AND Y.end_date = '2016'
),

b AS (
  SELECT DISTINCT
   chapter_code, chapter, section_code, section, para, subpara, chemical,
product, product_code,
   REPLACE(presentation,'GlucOsamine','prop-GlucOsamine') AS
presentation,REPLACE(presentation,' ','')
     AS presentation_no_spaces
   FROM ebmdatalab.hscic.bnf
   WHERE presentation NOT IN (SELECT presentation from
ebmdatalab.hscic.bnf_name_to_product_special_cases_helen)
   AND chapter_code <'18'),

a0 AS (
SELECT *,
      TRIM(bnf_7_char) AS bnf_7_char_trim,

SUBSTR(drug_name,1,IF(STRPOS(drug_name,'_')>0,STRPOS(drug_name,'_')-1,length(dr
ug_name)))
        AS drug_name_part,
      SUBSTR(drug_name,1,IF(STRPOS(drug_name,' ')>0,STRPOS(drug_name,'
')-1,length(drug_name)))
        AS drug_name_part_short,
      SUBSTR(chemical,1,IF(STRPOS(chemical,' ')>0,STRPOS(chemical,'
')-1,length(chemical)))
        AS chemical_short,
      REPLACE(drug_name,'GlucOsamine','prop-GlucOsamine') AS drug_name_a,

REPLACE(REPLACE(drug_name,'GlucOsamine','prop-GlucOsamine'),'Sulph','Sulf') AS
drug_name_b,
      REPLACE(new_chemical_name,'Sulph','Sulf') AS new_chemical_name_b
```

```
      FROM ebmdatalab.hscic.prescribing_pca_1998_2016_full_v2 a
      WHERE
IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,2),SUBSTR(bnf_7_char,1,2)) < '18'),

a1 AS (SELECT a0.*,
        z.new_bnf_code AS code_fuzzy,
        z.new_name AS drug_name_fuzzy,
        CONCAT( UPPER(substr(d.new_name,1,1)),
substr(D.new_name,2,LENGTH(D.new_name)-1) ) AS product_2013,
        E.alternative AS product_new_spelling,
        CONCAT( UPPER(substr(d1.new_name,1,1)),
substr(D1.new_name,2,LENGTH(D1.new_name)-1) ) AS chemical_2013,  -- note, this
capitalises the first letter only
        replace(a0.new_chemical_name_b,a0.chemical_short,D3.new_name) AS
chemical_2013b,
        replace(a0.drug_name_b,a0.drug_name_part,e.alternative) AS
converted_drug_name, -- incorporate new spellings into drug name
        replace(a0.drug_name_b,a0.drug_name_part,D.new_name) AS
converted_drug_name2,
        replace(a0.drug_name_b,a0.drug_name_part_short,D2.new_name) AS
converted_drug_name3,

SUBSTR(drug_name_b,1,IF(STRPOS(drug_name_b,'_')>0,STRPOS(drug_name_b,'_')-1,len
gth(drug_name_b))) AS drug_name_part_b

        FROM a0
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D ON
LOWER(drug_name_part) = D.old_name
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D1 ON
LOWER(a0.chemical) = D1.old_name
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D2 ON
LOWER(drug_name_part_short) = D2.old_name
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D3 ON
LOWER(a0.chemical_short) = D3.old_name
      LEFT JOIN temp E ON a0.drug_name_part = E.old_name AND
SUBSTR(a0.bnf_7_char_trim,1,4) = e.section_code
      LEFT JOIN ebmdatalab.hscic.pca_bnf_name_to_code_fuzzy_lookup z ON
A0.drug_name = z.old_name
      ),


--CAPITALISE WHERE NEEDED:
A2 AS (
SELECT *,
      CONCAT( UPPER(substr(converted_drug_name2,1,1)),
substr(converted_drug_name2,2,LENGTH(converted_drug_name2)-1) ) AS
converted_drug_name4,
```

```
      CONCAT( UPPER(substr(converted_drug_name3,1,1)),
substr(converted_drug_name3,2,LENGTH(converted_drug_name3)-1) ) AS
converted_drug_name5,
      CONCAT( UPPER(substr(chemical_2013b,1,1)),
substr(chemical_2013b,2,LENGTH(chemical_2013b)-1) ) AS chemical_2013_c
--capitalise chemical as well.
      FROM A1
      ),

--COALESCE TO FORM "FINAL" NAMES
a3 AS (
 SELECT *,
COALESCE(converted_drug_name,converted_drug_name4,converted_drug_name5,drug_nam
e_fuzzy,drug_name_b)
   AS drug_name_F,
 COALESCE(product_new_spelling,product_2013,drug_name_part_b) AS
drug_name_part_F,
 COALESCE(chemical_2013, chemical_2013_c, new_chemical_name_b) AS chemical_F
 from A2
 ),

 --add a drug name field without spaces:
 a4 AS (
 select *, REPLACE(drug_name_F,' ','')  as drug_name_F_no_spaces
 from A3
 ),


a AS  (
  SELECT
      x.bnf_7_char_trim AS bnf_code,
      x.drug_name,
      drug_name_F,
      COALESCE(spc.presentation,b.presentation,ba.presentation,
bb.presentation,bc.presentation, bd.presentation)
        AS current_bnf_name,
      COALESCE(spc.product_code,b.product_code,ba.product_code,
bb.product_code,bc.product_code, bd.product_code)
        AS current_bnf_code,
      drug_name_part,
      drug_name_part_F, -- use as product name if no other
      x.section,
      x.subpara,
      x.chemical AS Chemical_original,
      x.chemical_F AS Chemical,
      x.Year,
      SUM(x.owc2) AS OWC2,   -- prescribed generically but no generic available
      SUM(x.NIC) AS Cost,
```

```
      SUM(x.items) AS Items,
      SUM(x.quantity) AS Quantity
  FROM a4 x
    --AND A.Currently_in_BNF = 'N'
    LEFT JOIN ebmdatalab.hscic.bnf_name_to_product_special_cases_helen spc ON
upper(x.drug_name_F) = upper(spc.presentation)  -- look up original drug
details in current bnf (drugs matching more than one drug in bnf)
    LEFT JOIN b ON upper(x.drug_name_F) = upper(b.presentation)   -- use upper
to match up examples like this: "Pentasa Sr_Tab 250mg" and "Pentasa SR_Tab
250mg"
            AND SUBSTR(x.bnf_7_char_trim,1,4) = b.section_code   -- look up
original drug details in current bnf.

    LEFT JOIN b ba ON upper(x.drug_name_F) = upper(ba.presentation)
            AND SUBSTR(x.bnf_7_char_trim,1,4) != ba.section_code  -- check if
drug now only belongs in a different section but same chapter
            AND SUBSTR(x.bnf_7_char_trim,1,2) = ba.chapter_code
    LEFT JOIN b bb ON upper(x.drug_name_F) = upper(bb.presentation)
            AND SUBSTR(x.bnf_7_char_trim,1,2) != bb.chapter_code  -- check if
drug now only belongs in a different chapter
    LEFT JOIN b bc ON upper(x.drug_name) = upper(bc.presentation) AND
b.presentation IS NULL AND ba.presentation IS NULL AND bb.presentation IS NULL
--also check original in case new drug name didn't work e.g. nifedipin(e)
            AND SUBSTR(x.bnf_7_char_trim,1,4) = bc.section_code
    LEFT JOIN b bd ON x.drug_name_F_no_spaces = bd.presentation_no_spaces   --
match without spaces e.g. Terbut Sulf_Inha 250mcg (400 D) vs "(400D)"
            AND SUBSTR(x.bnf_7_char_trim,1,4) = bd.section_code  AND
b.presentation IS NULL AND ba.presentation IS NULL AND bb.presentation IS NULL
-- look up original drug details in current bnf.

  GROUP BY
      bnf_code, drug_name, drug_name_F, current_bnf_name, current_bnf_code,
      drug_name_part, drug_name_part_F, -- use as product name if no other
      section, subpara, Chemical_original, Chemical, Year
      )

SELECT
    a.bnf_code,
    a.current_bnf_code AS Product_code_updated,
    SUBSTR(COALESCE(a.current_bnf_code,b.product_code,a.bnf_code),1,2) AS
Chapter_code_current,
    SUBSTR(a.bnf_code,1,2) AS BNF_Chap_Code,
    COALESCE(b.chapter, ch.description) AS Chapter_Current,
    ch.description AS Chapter_original,
    SUBSTR(COALESCE(a.current_bnf_code,a.bnf_code),3,2) AS
Section_code_current,
    SUBSTR(bnf_code,3,2) AS BNF_Section_Code,
    COALESCE(b.section,se.description,a.section) AS Section_Current,
```

```
    a.section AS Section_Original,
    SUBSTR(COALESCE(a.current_bnf_code,b.product_code,a.bnf_code),5,2) AS
Para_code_current,
    COALESCE(b.para,pa.description) As Para_current,
    COALESCE(b.subpara,a.subpara) As Subpara_current,
    a.subpara AS Subpara_original,
    COALESCE(b.chemical,a.chemical) As Chemical_current,
    a.Chemical_original,
    COALESCE(b.product, a.drug_name_part_F) AS Product_current,
    current_bnf_name,
    a.drug_name,
    IF(b.product_code IS NULL,'N','Y') AS Currently_in_BNF,
    a.year,
    a.Items,
    a.owc2,
    a.Quantity,
    a.Cost

  FROM a
  LEFT JOIN ebmdatalab.hscic.bnf_vertical ch ON SUBSTR(a.bnf_code,1,2) =
ch.code
  LEFT JOIN ebmdatalab.hscic.bnf_vertical se ON SUBSTR(a.bnf_code,1,4) =
se.code
  LEFT JOIN ebmdatalab.hscic.bnf_vertical pa ON SUBSTR(a.bnf_code,1,6) =
pa.code
  LEFT JOIN b ON a.current_bnf_name = b.presentation
            AND a.current_bnf_code = b.product_code
Save results as ebmdatalab.tmp_eu.trends_from_pca_2016
```

B2b. Part 2

```
-- final pca data extraction (2016) part 2
-- distinct product-chemical combinations in current BNF:
WITH
 chem_p AS (
   SELECT DISTINCT product, product_code, chemical_code, chemical,
   count (distinct product_code) Over (partition by chemical_code, product)
      AS Dist_prods_with_same_name
   FROM ebmdatalab.hscic.bnf
   WHERE chapter_code <'18'
   ORDER BY Dist_prods_with_same_name, product),

-- find all drug_name_parts in PCA which have been mapped to a new chemical:
chem_0 AS (
  SELECT
```

```
SUBSTR(drug_name,1,IF(STRPOS(drug_name,'_')>0,STRPOS(drug_name,'_')-1,length(dr
ug_name)))
      AS drug_name_part,
  drug_name, section, old_chemical_name, new_chemical_name
  FROM ebmdatalab.hscic.pca_chemical_old_to_new_lookup_2016),

-- distinct *chemicals* in current BNF:
chem_a AS (
SELECT chemical,
    count(distinct chapter) AS Chapters,
    count(distinct section) AS Sections,
    count(distinct para) AS Paras,
    count(distinct chemical_code) AS Codes,
    min(chemical_code) AS min_code
  FROM ebmdatalab.hscic.bnf
  WHERE chapter_code < '18'
  GROUP BY chemical
  ORDER BY codes DESC, paras DESC, chemical),

-- for chemicals with multiple codes:
-- check whether each chemical code is the only one in its paragraph / section
/ chapter
chem_a1 AS
(SELECT DISTINCT
  a.chemical, b.chemical_code, a.paras, b.para_code, a.sections,b.section_code,
a.chapters, b.chapter_code,
  count(distinct b.chemical_code) over (partition by b.chemical,chapter_code)
    AS appearances_by_chapter,
  count(distinct b.chemical_code) over (partition by b.chemical,section_code)
    AS appearances_by_section,
  count(distinct b.chemical_code) over (partition by b.chemical,para_code)
    AS appearances_by_para
  FROM ebmdatalab.hscic.bnf b
  INNER JOIN chem_a a ON a.chemical = b.chemical and a.codes > 1
  WHERE b.chapter_code < '18'
  ORDER BY chemical ),


-- SELECT ALL CHEMICALS FROM BNF WHICH MAP TO A SINGLE PRODUCT
-- used in final step only
b AS (
  SELECT DISTINCT
   chapter_code, chapter, section_code, section, para_code, para, subpara_code,
subpara, chemical_code
   FROM ebmdatalab.hscic.bnf
   WHERE chapter_code <'18'),
```

```
t as (
 SELECT t.*,
 REPLACE(c2.new_chemical_name,'Streptokinase-Streptodornase','Streptokinase &
Streptodornase')
      AS new_chemical_name,
 CASE WHEN Product_current LIKE 'Levonelle%' THEN '0703050A0BC'   --
'Levonelle'
      WHEN Product_current LIKE 'Postinor%' THEN '0703050A0BB'    -- 'Postinor'
      WHEN t.drug_name LIKE 'Terbut%Sulph_Syr%' THEN '0301011V0AA'  -- 'Terbut
Sulf'
      WHEN t.drug_name LIKE 'Thalidomide%' AND Chapter_code_current = '05' THEN
'0501100J0AA' -- 'Thalidomide (Antileprotic)'
      WHEN Product_current LIKE 'Menoring 50' THEN '0702010G0BE'  -- 'Menoring
50'
      WHEN t.drug_name = 'Acetylcy_Eye Dps 10% (Old)' THEN '1108010C0AA' --
'Acetylcy (Eye)'
      WHEN t.drug_name = 'Abilify Maintena_Inj 400mg Vl + Dil' THEN
'0402020ADBB' -- 'Abilify Maintena'
      WHEN Product_current LIKE 'Melatonin%' THEN '0401010ADAA' -- 'Melatonin'
      WHEN Product_current LIKE 'Varidase%' THEN '1311070R0BB' -- 'Varidase'
      WHEN t.drug_name = 'Cocois_Scalp Oint' THEN '1305020V0BB' -- 'Cocois'
      WHEN t.drug_name = 'Levocarnitine_Oral Soln Paed 1.5g/5ml30%' THEN
'0908010C0AA' -- 'Levocarnitine'
      ELSE product_code_updated
      END AS product_code_updated_manual,
 CASE WHEN Product_current LIKE 'Levonelle%' THEN 'Levonelle'
      WHEN Product_current LIKE 'Postinor%' THEN 'Postinor'
      WHEN t.drug_name LIKE 'Terbut%Sulph_Syr%' THEN 'Terbut Sulf'
      WHEN t.drug_name LIKE 'Thalidomide%' AND Chapter_code_current = '05' THEN
'Thalidomide (Antileprotic)'
      WHEN Product_current LIKE 'Menoring 50' THEN 'Menoring 50'
      WHEN t.drug_name = 'Acetylcy_Eye Dps 10% (Old)' THEN 'Acetylcy (Eye)'
      WHEN t.drug_name = 'Abilify Maintena_Inj 400mg Vl + Dil' THEN 'Abilify
Maintena'
      WHEN Product_current LIKE 'Melatonin%' THEN 'Melatonin'
      WHEN Product_current LIKE 'Varidase%' THEN 'Varidase'
      WHEN t.drug_name = 'Cocois_Scalp Oint' THEN 'Cocois'
      WHEN t.drug_name = 'Levocarnitine_Oral Soln Paed 1.5g/5ml30%' THEN
'Levocarnitine'
      ELSE product_current
      END AS product_current_manual
 FROM ebmdatalab.helen.trends_from_pca_2016 t
 LEFT JOIN chem_0 c2 ON t.drug_name = c2.drug_name AND t.chemical_current =
c2.old_chemical_name AND SUBSTR(t.bnf_code,1,4) = c2.section
 ),
```

```
  A AS (
 SELECT T.*,
 COALESCE(chem_p.product,c2.product,product_current_manual) AS current_product,
    -- use this order in coalesce because we want to update/replace any
existing product names for which we now have a better one.
 COALESCE(product_code_updated_manual,chem_p.product_code,c2.product_code)
    AS current_product_code,

COALESCE(chem_p.chemical,c2.chemical,chem_a.chemical,chem_a1.chemical,c3.chemic
al)
    AS unique_chem,  -- chemicals currently in BNF (uniquely)

COALESCE(chem_p.chemical_code,c2.chemical_code,chem_a.min_code,chem_a1.chemical
_code,c3.min_code)
    AS unique_chem_code

 FROM  t
  -- link to BNF using whole Product name (note this will be drug_name_part)
--------------------
  -- chemical must match as well because product names are not always unique.
  LEFT JOIN chem_p ON t.product_current = chem_p.product
            AND t.Product_code_updated_manual IS NULL
            AND (UPPER(chem_p.chemical) = UPPER(Chemical_current)
                      OR  UPPER(chem_p.chemical) = UPPER(new_chemical_name))
            AND SUBSTR(chem_p.chemical_code,1,6) = SUBSTR(bnf_code,1,6)
                      -- some chemicals sit in multiple paras.
            AND chem_p.Dist_prods_with_same_name = 1

  -- try shortening Product names in BNF to match products in data (only if
whole name is not found) --
    LEFT JOIN chem_p c2 ON t.product_current =
SUBSTR(c2.product,1,length(t.product_current))
            AND t.Product_code_updated_manual IS NULL
            AND chem_p.product IS NULL
            AND UPPER(c2.chemical) IN (UPPER(Chemical_current),
UPPER(new_chemical_name))
            AND SUBSTR(c2.chemical_code,1,6) = SUBSTR(bnf_code,1,6)  --some
chems sit in multiple paras.
            AND chem_p.Dist_prods_with_same_name = 1

  -- link to BNF using "original" chemical name for chemicals which are unique
in BNF  ---------------
  LEFT JOIN chem_a ON UPPER(chem_a.chemical)  = UPPER(Chemical_current)
            AND chem_a.codes = 1 AND chem_p.chemical IS NULL
            AND t.Product_code_updated_manual IS NULL
  -- link to BNF using NEW chemical name for chemicals which are unique in BNF
--------------------
  LEFT JOIN chem_a c3 ON c3.chemical  = new_chemical_name
```

```
            AND c3.codes = 1 AND chem_a.chemical IS NULL
            AND chem_p.chemical IS NULL
            AND t.Product_code_updated_manual IS NULL
  -- link to BNF using NEW chemical name for chemicals which are NON-unique in
BNF  --------------
  -- provided that no chemical has been assigned in a previous join.
  -- first check same paragraph then section then chapter.
  LEFT JOIN chem_a1 ON chem_a1.chemical  = Chemical_current
            AND chem_p.chemical IS NULL
            AND chem_a.chemical IS NULL
            AND c3.chemical IS NULL
            AND t.Product_code_updated_manual IS NULL
            AND (
              (chem_a1.para_code = SUBSTR(bnf_code,1,6) AND
chem_a1.appearances_by_para = 1)
              OR (chem_a1.section_code = SUBSTR(bnf_code,1,4) AND
chem_a1.appearances_by_section = 1)
              OR (chem_a1.chapter_code = SUBSTR(bnf_code,1,2) AND
chem_a1.appearances_by_chapter = 1)
                    )
 ORDER BY drug_name,year ),

 u AS (
 select bnf_code,
 Chapter_code_current,   BNF_Chap_Code,     Chapter_Current, Chapter_original,
 Section_code_current,   BNF_Section_Code,  Section_Current,   Section_Original,
 Para_code_current,      Para_current,
 Subpara_current,  Subpara_original,
 COALESCE(unique_chem_code, SUBSTR(Product_code_updated,1,9),
SUBSTR(current_Product_code,1,9))
   AS chem_code_today, --chemical code
 Chemical_original,
 COALESCE(unique_chem,Chemical_current) AS chem_today,
 COALESCE(Product_code_updated, current_product_code)    AS prod_code_today,
 COALESCE(current_product,     Product_current) AS prod_today,
   -- note this is opposite way around to code because we want to replace the
previous name
   -- but there may not be a code.
 current_bnf_name, drug_name,
 Currently_in_BNF,
 year,  Items,  owc2,  Quantity,  Cost
 FROM a
 ORDER BY drug_name, year)


SELECT
    bnf_code,
    COALESCE(b.chapter_code,Chapter_code_current) AS Chapter_code_current,
```

```
    BNF_Chap_Code,
    COALESCE(b.chapter,Chapter_Current) AS Chapter_Current,
    Chapter_original,
    COALESCE(SUBSTR(b.section_code,3,2),Section_code_current) AS
Section_code_current,
    BNF_Section_Code,
    COALESCE(b.section,Section_Current) AS Section_current,
    Section_Original,
    COALESCE(SUBSTR(b.para_code,5,2),Para_code_current) AS Para_code_current,
    COALESCE(b.para,Para_current) AS Para_current,
    COALESCE(b.subpara,Subpara_current) AS Subpara_current,
    Subpara_original,
    chem_code_today AS Chemical_code_current,
    Chemical_original,
    chem_today AS Chemical_current,
    prod_code_today AS Prod_code_current,
    prod_today AS product_current,
    current_bnf_name, drug_name,
    Currently_in_BNF, u.year,  Items,  owc2,  Quantity,  Cost,
    -- add calculated fields:
    1000*items/pop.Population AS ItemsPer1000,
    1000*quantity/pop.Population AS QuantityPer1000,
    Inf.Multiplier_2016*cost AS Infl_corr_Cost,
    1000*Inf.Multiplier_2016*cost/pop.Population AS Infl_corr_Cost_per1000,
    IEEE_DIVIDE(Inf.Multiplier_2016*Cost, Items) AS Infl_corr_CostPerItem,
    1000*owc2/pop.Population AS Owc2Per1000

 FROM U
 LEFT JOIN b ON u.chem_code_today = b.chemical_code
 LEFT JOIN ebmdatalab.ONS.england_midyear_population pop ON u.Year = pop.Year
 LEFT JOIN ebmdatalab.ONS.inflation_cpi inf ON u.Year = inf.Year
 --WHERE LENGTH(chem_code_today) =8
```

BMJ Open

OpenPrescribing: Normalised Data and Software Tool to Research Trends in English NHS Primary Care Prescribing 1998-2016

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019921.R2
Article Type:	Research
Date Submitted by the Author:	11-Dec-2017
Complete List of Authors:	Curtis, Helen; University of Oxford Department of Primary Care Health Sciences Goldacre, Ben; University of Oxford, Primary Care Health Sciences
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	General practice / Family practice
Keywords:	Prescribing, National Health Service (NHS), PRIMARY CARE

OpenPrescribing: Normalised Data and Software Tool to Research Trends in English NHS Primary Care Prescribing 1998-2016

Helen J Curtis[1], Ben Goldacre[1,*]

*Corresponding author: ben.goldacre@phc.ox.ac.uk

[1]Evidence Based Medicine DataLab
Centre for Evidence Based Medicine
Department of Primary Care Health Sciences
University of Oxford
Radcliffe Observatory Quarter
Woodstock Road
Oxford OX2 6GG

Word Count: 3,915

ABSTRACT

Objectives: We aimed to compile and normalise England's national prescribing data for 1998-2016 to facilitate research on long-term time trends, and create an open data exploration tool for wider use.

Design: We compiled data from each individual year's national statistical publications and normalised them by mapping each drug to its current classification within the national formulary where possible. We created a freely accessible, interactive web tool to allow anyone to interact with the processed data.

Setting and Participants: We downloaded all available annual prescription cost analysis datasets, which include cost and quantity for all prescription items dispensed in the community in England. Medical devices and appliances were excluded.

Primary and secondary outcome measures: We measured the extent of normalisation of data and aimed to produce a functioning accessible analysis tool.

Results: All data were imported successfully. 87.5% of drugs were matched exactly on name to the current formulary, and a further 6.5% to similar drug names. All drugs in core clinical Chapters were reconciled to their current location in the data schema, with only 1.26% of drugs not assigned a current chemical code. We created an openly accessible interactive tool to facilitate wider use of these data.

Conclusions: Publicly available data can be made accessible through interactive online tools, to help researchers and policymakers explore time trends in prescribing.

Strengths and limitations of this study

- We processed publicly-available annual data for the whole of England's community dispensing - not a sample.
- We corrected for population size, inflation, and (where possible) drugs changing name and/or classification over time.
- We produced a free, openly accessible tool for wider use, displaying trends in items, cost, price-per-item and quantity-per-item for each product for 1998-2016, which can be updated annually.
- The tool is limited to product-level data, not individual presentations, and wide-scale correction for dosage was not possible.
- Users can also download our normalised dataset in order to carry out their own analyses.

Abbreviations

BNF - British National Formulary
BSA - NHS Business Services Authority
CNS - Central nervous system
CPI - Consumer Price Index
CPRD - Clinical Practice Research Datalink
GIS - Gastrointestinal system
NHS - National Health Service
NIC - Net Ingredient Cost
ONS - Office of National Statistics
PCA - Prescription Cost Analysis

INTRODUCTION

In 2016, NHS prescribing in England cost £9.20bn [1], approximately 9% of the annual NHS budget [2]. Prescribing behaviour is expected to respond within the dynamic system of evidence-based medicine, through changing patterns of disease, innovation in medical treatments, and new evidence. Monitoring long-term time trends in prescribing is therefore useful to observe changes in practice, to provide a form of feedback to ensure there are no unexpected or undesirable changes, and to facilitate tracking and forecasting of costs.

NHS Digital publish monthly and annual prescribing datasets from the NHS Business Services Authority (NHSBSA), along with static reports on prescribing trends. However this does not allow readers to interrogate topics of interest in detail, and the large datasets can be complex to manage. We provide a service at OpenPrescribing that facilitates exploration of outliers and trends for individual general practices in NHS England, which has provided over 250,000 analyses to 50,000 users over the past year. The detailed dataset that drives this service (running to over 10m rows a month) is only available from 2010 onwards. The annual Prescription Cost Analysis (PCA) data, aggregated nationally (with no data on individual practice), and by year (with no data on prescribing changes each month) are available back to 1998. These data are freely accessible, but consist of individual files for each year of prescribing, which cannot be straightforwardly combined, and therefore do not facilitate interrogation of time trends. Additionally, identifiers for individual drugs may change name, or location within the British National Formulary (BNF), over time, making simple compilation of the data impossible.

The value of PCA data is indicated by the numerous previous studies using it to assess prescribing trends [3–5] or to detect changes in response to guidelines or safety alerts [6–8]. These studies have been focused on data for a small number of drugs, manually aggregated for each bespoke analysis; furthermore, given publishing delays for academic manuscripts the data are commonly very delayed, and readers cannot easily place the findings in context of current clinical practice or expenditure.

We therefore set out to aggregate all available PCA data into a single data frame for longitudinal analysis of trends, in a service that could be easily updated; to generate an interactive online service where any user can explore and monitor time trends in prescribing using the latest available data; and to share all resources for re-use by others as open data.

METHODS

Data sources

Every available Prescription Cost Analysis (PCA) annual dataset was downloaded from NHS Digital or National Archives, covering calendar years 1998 to 2016 [9].

Data structure

Each annual PCA dataset includes all items dispensed in England by pharmacy/appliance contractors, dispensing doctors, and items personally administered by doctors, whether or not they were *prescribed* in England or other parts of the UK. Items dispensed in other settings (prisons, hospitals and private prescriptions) are excluded. Prior to 2010, the data were rounded to the nearest 100 and excluded drugs with fewer than 50 items prescribed, accounting for 0.01% of total items [10]. Definitions of key terms used in the PCA data (and NHS primary care prescribing data more generally) are given in Box 1 and a full glossary of terms is available [11].

Box 1. Glossary of prescribing data terminology.

Example presentation: Tradorec XL Tablets 300mg

Drug Name	BNF Chemical Name	BNF Section Name	BNF Sub Paragraph Name	Items	Quantity	NIC (£)
Tradorec XL_Tab 300mg	Tramadol Hydrochloride	Analgesics	Opioid Analgesics	6,374	324,167	152,358

- The *drug name* describes the full *presentation* of the drug, i.e. the formulation and strength as well as the drug's brand or generic (*product*) name.
- The *chemical name* is the standard registered name for the active constituent of the medicine. It is not always an individual chemical: examples include "Paracetamol Combined Preparations" and "Paracetamol & Caffeine".
- Numerical codes representing Chapter, Section, Paragraph and Sub-paragraph are also supplied. These represent only the first seven characters of each drug's unique 15-character BNF code - see Box 2.
- *Items* are functionally equivalent to prescriptions; they do not take into account the quantity (number of boxes/bottles etc.) dispensed to the same person. *Items* may vary in the *quantity* prescribed.
- *Quantity* represents the quantity of a drug dispensed, with units of measurement (units/tablets/grammes/millilitres etc.) dependent upon its formulation.
- *Net Ingredient Cost* (NIC) represents the basic price of the medicine, i.e. the Drug Tariff price, or, if not listed, the price published by the manufacturer or supplier. NIC may be subject to further charges and/or discounts. Patients who are eligible contribute a fixed fee towards each prescription charge, but this only applies to a minority of items and it is not possible to identify which items in this dataset.

Every drug presentation (i.e. each formulation, dose and product combination) is described by a unique drug name, and has a unique 15-digit structured British National Formulary (BNF) code, an example of which is given in Box 2. The BNF contains an entry for every product available to be prescribed in Britain, including medicinal products, dietary supplements, complementary therapies and physical appliances such as bandages. The hierarchical BNF codes imply a data schema as follows: each *presentation* of a drug has a *product* name, which may be either a brand name or the generic *chemical* name; as such, each product can be mapped to a chemical. Each chemical is a member of a Paragraph in the BNF (some of which are divided into Sub-paragraphs, which themselves often approximate to a class of drugs). Each Paragraph belongs to a Section, which is in turn a member of a Chapter (often approximating to a system of the body, such as "Cardiovascular").

Box 2. BNF Code Structure.

Example presentation: Tradorec XL Tablets 300mg

Chapter	Section	Paragraph	Sub-paragraph	Chemical	Product	Presentation	Generic Equivalent
04	07	02	0	40	BI	AC	AM
Central Nervous System	Analgesics	Opioid Analgesics	Opioid Analgesics	Tramadol Hydrochloride	Tradorec	Tradorec XL_Tab 300mg	*

*Generic equivalent allows matching with the strength and formulation (presentation) of the generic product (which will always have product code 'AA').
For generic presentations the *product* name will match the *chemical* name (but sometimes with a different abbreviation, e.g. "Tramadol HCl").

In the PCA data, only the first seven characters of the BNF code for each drug are supplied, rather than the full BNF code. Therefore,the "drug name" is the only source of information on the formulation and dose ; however, from this the BNF code can usually be imputed, but this becomes increasingly difficult for older drugs no longer listed in the BNF. Although each drug's chemical name is also supplied, chemicals are not all unique (e.g. "Other Preparations"); names may change their spelling over time; and chemicals may move between Paragraphs, Sections and Chapters. Indeed, classifications at any level of the hierarchy can be subject to renaming, spelling change, subdivision, reorganisation and removal.

Data management, aggregation, and cleaning

All data were grouped by drug name, combining those differing only by standard quantity unit (SQU). Ultimately, following cleaning, data were grouped to product level. Medical devices/appliances and any other items Chapter numbers above 15 were excluded.

A key user-need was to explore prescribing trends for individual members of a class of drugs over time. This required all data to be normalised, with each individual drug consistently appearing in the correct location in the data schema; i.e. all individual presentations of a chemical all mapped under that chemical; and all chemicals mapped under the correct Sub-paragraph/Paragraph (often similar to drug class) of the BNF. To achieve this consistency, we aimed to map each drug to its current position in the latest BNF dictionary, up to the level of its 11-character "product" code, through an incremental process. This is summarised below and in Figure 1.

Lacking the full BNF code, we attempted to match each drug name to a current BNF presentation. Those without an exact match (e.g. formulation variants no longer available) could sometimes be matched to a similar BNF presentation name, e.g. by finding a similar formulation or using the "fuzzy" lookup add-on for Excel and validated manually [12]. Other drug names could only be matched up to current BNF codes by using their product or chemical names. Matching at each stage was improved by disregarding capitalisation, or spacing and spelling changes (e.g. Sulphur/Sulfur); these include changes identified within the data and those occurring when many old British spellings (the "British Approved Name") were replaced with international standard names (the "Recommended International Non-Proprietary Name") [13]. Remaining drug names in the most-prescribed Chapters (1-6 and 10) were matched to current drug names manually (for example, resolving non-matches due to rearrangement of word order); any others kept original chemical name, and a proxy product name was derived from the drug name field. Full methodology for this matching process is available in our technical documentation online [14] and in Supplementary Material.

We measured the extent of normalisation of drug names and classifications, and present summary statistics on these.

Normalisation for inflation and population

Prescribing costs were corrected for inflation using the UK's annual consumer price index (CPI) figures, normalised to 2016 [15]. Number of items prescribed and costs were divided by the population each year to calculate values per thousand population, based upon mid-

year population estimates for England only [16]. We also supply the original number of items and cost in our output.

Interactive Analysis Tool

Having generated a normalised dataset, and a method for updating it, we then set out to implement a free, interactive online data analysis tool where any user can visually explore time trends in prescribing. This was built using Tableau Public, a freely accessible interactive data presentation platform which permits rapid prototyping; however other front-ends onto the same underlying datasets could also be implemented using open source tools such as Shiny [17] or in Python libraries such as d3 [18] with more software engineer resource. Our user-needs for the tool were as follows: the ability to display trends in items and cost, normalised for total population change and inflation; and also to calculate the average cost per item and quantity per item for each product.

Having delivered the tool, we used it to generate trends data and graphs for a range of clinical areas where prescribing trends have been previously studied and published, to demonstrate the ability of our tool to replicate and extend these works.

Data and code

The full compiled and processed PCA datasets are available online via FigShare [19], SQL code is in Supplementary File, and the Trends tool (Tableau workbook) is available to use via https://OpenPrescribing.net/pca.

RESULTS

Data compilation and overall prescribing trends

All data were successfully imported. There were 169,100 lines of data in the compiled 1998-2016 dataset (Chapters 1-15) and 169,038 in the processed data, the reduction caused by aggregation of a small number of drugs available in multiple formulations despite having identical names. Total items (14.8 billion), cost (£136.9 billion) and distinct drug names (22,496) remained consistent before and after data processing (Table 1, S1). The inclusion of low volume prescribing in the published datasets from 2010 caused a substantial rise in the number of distinct drugs per year, but not items or cost (Table 1). As can be seen from Table 1, the inflation-corrected cost in 2016 UK sterling equivalent for all prescribing in NHS England primary care rose from £6.3bn in 1998 to £10.1bn in 2004, but then decreased to £8.3bn in 2016. Items per 1,000 population (correcting for crude population growth) has grown from 10,180 in 1998 to 19,196 in 2016, on average increasing by 3.6% per year. The discrepancy between items and cost is caused by variation in the choice of drugs being prescribed and their individual prices, for example switching from branded drugs to cheaper generic versions after patent expiry.

Table 1. Summary of processed PCA data by year (drugs in Chapters 1-15 only). 'Drug Name' is the field describing the presentation of each drug, i.e. its formulation, dose and product name. Costs represent Net Ingredient Cost (NIC, see Box 1). 'Change' is the year-on-year change.

Year	Distinct count of Drug Name	Items	Items per 1000		Cost	Inflation-Corrected Cost	Inflation-Corrected Cost per 1000	
	n	n	n	change (%)	£	2016 £	2016 £	change (%)
1998	6,338	497.0M	10,180		£4,440M	£6,280M	£128,626	
1999	6,587	513.4M	10,471	+2.9%	£5,011M	£6,999M	£142,746	+11.0%
2000	6,613	535.1M	10,868	+3.8%	£5,284M	£7,318M	£148,648	+4.1%
2001	6,754	569.2M	11,510	+5.9%	£5,784M	£7,914M	£160,049	+7.7%
2002	6,834	598.6M	12,050	+4.7%	£6,487M	£8,768M	£176,491	+10.3%
2003	6,893	630.3M	12,625	+4.8%	£7,113M	£9,488M	£190,035	+7.7%
2004	6,912	666.0M	13,268	+5.1%	£7,645M	£10,063M	£200,482	+5.5%
2005	6,907	698.8M	13,808	+4.1%	£7,452M	£9,609M	£189,875	-5.3%
2006	6,810	728.4M	14,292	+3.5%	£7,660M	£9,655M	£189,436	-0.2%
2007	7,056	771.8M	15,022	+5.1%	£7,810M	£9,614M	£187,112	-1.2%
2008	7,202	816.7M	15,762	+4.9%	£7,716M	£9,174M	£177,047	-5.4%
2009	7,401	859.2M	16,461	+4.4%	£7,892M	£9,176M	£175,805	-0.7%
2010	11,703	898.4M	17,065	+3.7%	£8,162M	£9,193M	£174,636	-0.7%
2011	11,751	931.6M	17,541	+2.8%	£8,101M	£8,734M	£164,457	-5.8%
2012	12,207	968.9M	18,112	+3.3%	£7,802M	£8,176M	£152,836	-7.1%
2013	12,318	996.2M	18,494	+2.1%	£7,846M	£8,022M	£148,920	-2.6%
2014	12,576	1,027.0M	18,908	+2.2%	£8,022M	£8,078M	£148,718	-0.1%
2015	12,875	1,043.5M	19,046	+0.7%	£8,403M	£8,461M	£154,444	+3.9%
2016	13,285	1,060.9M	19,196	+0.8%	£8,284M	£8,284M	£149,892	-2.9%

Total	**22,496**	**14,810.9M**	**284,680**	**+3.6%**	**£136,914M**	**£163,006M**	**£3,160,254**	**1.0%**

Data Normalisation

Data were normalised using the methods described above. Of the distinct drug names in the data, 87.5% were matched exactly to a current BNF name, and a further 6.5% matched approximately (Table 2). Name changes are particularly prevalent in Chapter 3 (Respiratory), due mainly to the addition in 2004 of a space when a number of doses is given, as is common for inhalers, e.g. “Salbutamol_Inha 100mcg (200D)” became “Salbutamol_Inha 100mcg (200 D)”.

Table 2. Number and percentage of drug names subject to changes within 1998-2016 PCA data when compared to the current BNF, by Chapter. These include changes in word order, spacing, capitalisation, abbreviation, punctuation (e.g. “Califig_(California Syr Of Figs)”/“Califig_California Syr Of Figs”), spelling (e.g. “Sulphate”/“Sulfate”), brand name (e.g. “Laxoberal_Liq”/ “Dulcolax Pico_Liq”) and formulation (e.g. “Castor Oil_”/“Castor Oil_Liq”). The total count of drug names is reduced compared to Table 1 because the same drugs can appear over multiple years but only rarely in multiple Chapters.

Current Chapter code (name)	Name/spelling change		No change		No match		Grand Total
	n	%	n	%	n	%	n
1 (Gastro–intestinal system)		0.0%	1,041	99.1%	9	0.9%	**1,050**
2 (Cardiovascular system)	177	7.2%	2,278	92.5%	8	0.3%	**2,463**
3 (Respiratory system)	238	20.8%	893	78.0%	14	1.2%	**1,145**
4 (Central nervous system)	341	8.8%	3,535	90.7%	21	0.5%	**3,897**
5 (Infections)	119	9.1%	1,184	90.7%	3	0.2%	**1,306**
6 (Endocrine system)	203	13.2%	1,320	86.0%	11	0.7%	**1,534**
7 (Obstetrics, gynaecology and urinary–tract disorders)	4	0.7%	508	90.7%	48	8.6%	**560**
8 (Malignant disease and immunosuppression)	2	0.3%	538	91.8%	46	7.8%	**586**
9 (Nutrition and blood)	210	4.0%	4,281	81.3%	776	14.7%	**5,267**
10 (Musculoskeletal and joint diseases)	89	8.5%	956	90.8%	8	0.8%	**1,053**
11 (Eye)	18	3.0%	532	87.2%	60	9.8%	**610**
12 (Ear, nose and oropharynx)	16	3.4%	373	78.5%	86	18.1%	**475**
13 (Skin)	22	1.1%	1,772	89.5%	186	9.4%	**1,980**
14 (Immunological products and vaccines)	7	3.0%	198	85.0%	28	12.0%	**233**

15 (Anaesthesia)	15	5.2%	229	78.7%	47	16.2%	**291**
Grand Total	1,461	6.5%	19,638	87.5%	1,351	6.0%	**22,450**

Of the distinct drug names (23,275, taking into account some drugs having multiple BNF classifications), over 91% could be matched to a current product in the BNF, with no change in code (Table 3). Less than 5% could not be matched to a current product and/or chemical code, under 1% of items prescribed. These drugs were assigned proxy product names (derived from their drug name) so that all data could be presented visually, and for those not matched to a current chemical, the original chemical name was used (mostly "Other Preparations"). However, normalisation was focused on seven of the most prescribed Chapters with the greatest medical interest (1-6 and 10). The normalisation of drugs in other Chapters could therefore potentially be improved. In particular, Chapters 9 (Nutrition) and 13 (Skin) have substantial levels of prescribing, but are complex, containing many different drug names and non-drug products such as topical applications and dietary supplements. Other groups with a particular interest in nutrition or dermatology may wish to expand our work on manual matching: we would be happy to incorporate such amendments into our dataset. Many of the code changes and non-matches have diminished over time, as expected (Table S2).

Code changes and normalisation outputs are described in Table 3. Headers indicate the highest level in the BNF hierarchy at which drugs have been subject to code changes, e.g. "Section" indicates drug names which have not changed Chapter but have moved Section. "No product match" indicates drug names matched to a chemical (9-character BNF) but with no current matching product (11-character). "No chemical match" indicates drug names matched neither to a chemical nor product. The total count of drug names increases when separated by Chapter because four drug names currently exist in two different Chapters.

Table 3. Summary of drug code changes within the 1998-2016 prescribing datasets, also separated by (current) Chapter. Chapter names can be found in Table 2.

	BNF code change							
	Chapter	**Section**	**Paragraph**	**Sub-paragraph**	**No change**	**No product match**	**No chemical match**	**Grand Total**
Distinct count of Drug Name	94	52	560	203	21,258	815	293	**23,275**
% of Drugs	0.40%	0.22%	2.41%	0.87%	91.33%	3.50%	1.26%	**100%**
% of Items	0.04%	0.01%	0.84%	0.51%	97.67%	0.84%	0.10%	**100%**
Distinct count of Drug Name by Current Chapter								
Current Chapter number	**Chapter**	**Section**	**Paragraph**	**Sub-paragraph**	**No change**	**No product**	**No chemical**	**Grand Total**

						match	match	
1		4	109		1,146	6		**1,265**
2	5	3	4	19	2,446	7		**2,484**
3	1	6	19		1,132	10		**1,168**
4	42	13	173		3,831	9		**4,068**
5	1		161	116	1,241	2		**1,521**
6	4	1	2	2	1,521	10		**1,540**
7	5		12		528	28		**573**
8	1		4		579	6	1	**591**
9	1	7	6	54	4,510	507	245	**5,330**
10	27				1,041	9		**1,077**
11	1	5	5		571	28	6	**616**
12					420	44	11	**475**
13	5	13	28	12	1,847	95	30	**2,030**
14			37		195	17		**249**
15	1				254	37		**292**
Grand Total	**94**	**52**	**560**	**203**	**21,262**	**815**	**293**	**23,279**

Interactive Data Analysis Tool

We created a tool which allows anyone to explore the prescribing data, available directly at https://openprescribing.net/pca. Users can search by chemical, Paragraph, Section or Chapter to view time trends in items and costs on stacked charts, where both the overall trends and the relative contribution from each product/chemical can be seen. The cost per item and quantity per item for each product are also shown, which can assist in interpretation of trends in some cases. However, these calculations carry a “use with caution” note, as items may represent different pack sizes, and quantities cannot be reliably summed across preparations because of different strengths and formulations. The page features an accompanying video walk-through demonstrating the tool.

The tool can be used to facilitate novel research into time trends, and factors associated with changes in practice such as publication of guidelines or evidence landmarks, or changes in price. It can replicate and extend the main findings of previous papers which researched trends for different clinical areas using PCA data. For example, the antipsychotic drug switches which occurred in England following a licence restriction [20] can be replicated in the tool and the trends extended to the latest data (Figure 2a). This also shows the dramatic reductions in cost that followed the expiry of patents for risperidone and olanzapine. We also replicate antidepressant prescribing trends, previously reported up to 2010 [5], and show that how the overall use of these antidepressants has continued to rise, in particular sertraline (Figure 2b). We also replicate findings on the rise of thyroid hormones [3] and

testosterone [4], where we show that prescribing of these drugs continued to rise, with a disproportionate increase in cost (Figure S1a-b). We are using this dataset and tool in our academic papers on trends and variation in NHS prescribing; we encourage others to use our dataset and tool in their own work.

The tool can also be used to complement studies performed in more detailed prescribing data such as the Clinical Practice Research Datalink (CPRD), by giving the full national picture, and giving more longitudinal data that updates with new data releases. For example, several previous publications have reported on patterns of prescribing of smoking cessation medication in The Health Improvement Network (THIN) database [21–23]. This included reporting of a possible decline in prescribing despite increased incentives for GPs introduced in 2012. We can confirm this decline and show that it continued beyond 2013 (Figure 2c). We also show that the slow decline in quinine usage following safety alerts in 2010 [24] has continued at a similar pace (Figure S1c). CPRD data contain individual patient records and can therefore be used to assess detailed questions about treatments in specific cohorts of patients. However many labour-intensive CPRD analyses have been conducted to interrogate simple broad prescribing trends which could more straightforwardly be conducted using aggregated and normalised national data, with greater coverage of years and total population. Furthermore, for analyses interrogating national trends and responses to guidelines, in many cases a prescribing change which can only be detected in individual patients' records, and cannot be detected in national data, may not be relevant in terms of population health or the health service.

Additional tabs in the tool allow discovery of higher-level trends, including Chapter and Section trends, Sections ranked by items/cost for any selected year, calculation of the change in items/cost for each Section between any selected year to the latest year, and the top 20 Paragraphs by items and cost. The Chapter trends page, for example, shows that much of the decline in prescribing costs since the peak in 2004 (Table 1) is attributable to a drop in the cost of cardiovascular drugs (Figure 3a), and the Section trends page further shows that lipid-regulating drugs (Section 2.12) and Drugs for Hypertension and Heart Failure (Section 2.5) experienced the largest cost reductions at that time (Figure 3b).

DISCUSSION

Summary

It was possible to aggregate all PCA data from 1998-2016 and normalise for most changes in drug names and classifications. Only 87.5% of drug names matched exactly to a current

BNF name and 8.7% had undergone some change in classification; however all drugs in core clinical Chapters were reconciled to their current location in the data schema. We generated an interactive online service where any user can explore time trends in prescribing broken down by product, chemical, Paragraph, Section and Chapter; this openly accessible interactive data analysis tool provides overviews and insights comparable to previous labour-intensive bespoke data analysis research projects.

Strengths and weaknesses of this study

Our tool covers the data for the whole of England's community dispensing, not a sample. We are surprised to note that this is the first project aiming to aggregate long-term trends across the entire prescribing dataset, and provide an openly accessible tool for wider use. Many drugs changed name and/or classification over time, but valid chemicals were successfully assigned to all items in Chapters 1-6 and 10, and product names were derived for every drug, allowing maximum consistency in trends analysis. The tool is limited to product-level data due to the wide number of different presentations available.

We used items to measure prescribing volume. Quantity is generally more complex for making comparisons as there is wide variation caused by the units, which may be the number of pills or millilitres, units (such as inhalers containing multiple doses) or other unit measure. Converting quantities to approximate daily dose sizes (such as Defined Daily Doses, DDDs) is possible, but the conversion tables available are not sufficiently comprehensive to allow this across the entire dataset, and this would be even more difficult for discontinued drugs. We therefore rejected this option in favour of being able to publish a complete dataset. However, users wishing to analyse data by daily doses can download our BNF-normalised dataset in order to apply these calculations. Using items also has limitations, as it does not take into account number of packs prescribed per prescription, pack size or dosage. We are launching this tool publicly and will monitor user volume and user-feedback: if appropriate we will improve the tool by replicating and expanding it using bespoke software as per our other data analysis tools on OpenPrescribing.net for exploring variation in prescribing at CCG and individual practice level. We will update the tool annually, dependent upon continuing funds for the OpenPrescribing project.

Findings in context of other research

Long-term trends in prescribing have previously been reported on a wide variety of clinical areas, using PCA data as well as other sources [3–5,20,22,23]. These are static, not updated, and rapidly out of date. Although using CPRD allows a more detailed analysis and investigation of patient factors associated with prescribing, it takes a great deal of

preparation and time to complete. Our tool can replicate some trends found in CPRD, and so may provide a useful tool for preliminary investigation of trends. It can also help to confirm whether findings from regional datasets of rich individual patient data (IPD) from electronic health records sources are representative of the national picture, while avoiding repeated work and replication in new IPD datasets. In our related publications on variation and trends in specific disease areas we report comparisons between trends in PCA data, and trends from other more labour-intensive sources such as CPRD, in more detail.

The UK government produces a 10-year trends document following the annual PCA data release, containing an overall summary of high-level trends and a brief breakdown of six interesting topics with the greatest level or change in prescribed items and cost [25]. However, the reported topics are few in number, chosen by NHS Digital, restricted to ten years of data, do not correct for inflation or population growth, are not easily discoverable by subject specialists, and readers are not able to interrogate their own topics of interest in detail. From 2016, the compiled datasets were also made available so users may conduct their own exploration of the data, but, without drug names or categories being normalised, this is little better than the raw data, which we have processed into a normalised longitudinal dataset.

Policy implications and future research

Published papers can provide a useful and detailed insight into prescribing trends [3–5], but give a single snapshot which may quickly become out of date. Our tool facilitates ongoing monitoring by researchers and policymakers to assess prescribing changes in any area of concern or clinical interest they have identified; and permits interactive exploration of detailed issues in the data, such as individual presentations of chemicals, by any interested user. As part of our OpenPrescribing work we are using prescribing data to investigate adherence to guidelines and changes in practice in various clinical areas, to detect anomalous changes in individual practices relative to national trends in order to send practices alerts, and to identify cost saving opportunities. We have produced various manuscripts using the longitudinal data presented here as part of a range of data sources to describe variation in prescribing in primary care. We are happy to collaborate with other teams of clinicians and academics; we also release our underlying dataset and code as open data for re-use with citation.

Conclusions

Long-term trends in prescribing are interesting for a number of applications. While previous work on prescribing data has focused on static, manual analysis of a small number of drugs,

modern data science approaches make it possible to create interactive services that allow clinicians, healthcare commissioners, policy makers, academics and any other interested party to interrogate and monitor prescribing trends for any combination of chemicals, to identify anomalies or signals of concern, and predict spending. We have delivered this using a combination of open data and freely accessible online tools.

ACKNOWLEDGEMENTS

We are grateful to Seb Bacon for maintaining databases and general assistance, and to Richard Croker for pharmaceutical advice and fuzzy matching in Excel.

FUNDING

OpenPrescribing is funded by: NIHR Biomedical Research Centre Oxford, Health Foundation (Ref 7599), NIHR SPCR (Ref 327). No specific funding was sought for this project. The authors' funders had no involvement in the study design or the decision to submit.

CONFLICT OF INTEREST

All authors have completed the Unified Competing Interest form (available on request from the corresponding author) and declare: BG has received research funding from the Laura and John Arnold Foundation, the Wellcome Trust, the NHS NIHR School of Primary Care, the Health Foundation, NHS England, NIHR Biomedical Research Centre Oxford, and the WHO; he also receives personal income from speaking and writing for lay audiences on the misuse of science. HC is employed on BG's OpenPrescribing grants.

CONTRIBUTORSHIP STATEMENT

BG conceived and supervised the project, HC designed the methods, conducted the analysis, interpreted the findings, extracted and processed the data in BigQuery, Excel and Tableau with input from BG. HC and BG wrote the paper. All authors contributed to and approved the final manuscript. BG is guarantor.

REFERENCE LIST

1 NHS-Digital. National Statistics Prescription Cost Analysis, England - 2016. 2016.http://www.content.digital.nhs.uk/searchcatalogue (accessed 28 Sep 2017).

2 NHS-Choices. About the National Health Service (NHS) in England - NHS Choices. 2016.http://www.nhs.uk (accessed 28 Sep 2017).

3 Mitchell AL, Hickey B, Hickey JL, *et al.* Trends in thyroid hormone prescribing and consumption in the UK. *BMC Public Health* 2009;**9**:132.

4 Gan EH, Pattman S, H S Pearce S, *et al.* A UK epidemic of testosterone prescribing, 2001-2010. *Clin Endocrinol* 2013;**79**:564–70.

5 Ilyas S, Moncrieff J. Trends in prescriptions and costs of drugs for mental disorders in England, 1998–2010. *Br J Psychiatry* 2012;**200**:393–8.

6 Dietrich ES. Effects of the National Institute for Health and Clinical Excellence's technology appraisals on prescribing and net ingredient costs of drugs in the National Health Service in England. *Int J Technol Assess Health Care* 2009;**25**:262–71.

7 Connor AJ, Fraser SG. Glaucoma prescribing trends in England 2000 to 2012. *Eye* 2014;**28**:863–9.

8 Fitzpatrick RW, Pate RG. Assessing the impact of NICE guidance on the prescribing of hormonal treatments of breast cancer in England. *J Eval Clin Pract* 2015;**21**:759–61.

9 NHS-Digital. Prescribing Data. 2017.https://digital.nhs.uk/article/4214/Prescribing (accessed 21 Sep 2017).

10 NHS-Digital. Prescription Cost Analysis - England, 2009 [NS]. 2012.https://digital.nhs.uk/catalogue/PUB01414 (accessed 11 Sep 2017).

11 NHS-Digital. Prescription Cost Analysis Glossary 2015. pres-cost-anal-eng-2015-apx.pdf. 2016.http://content.digital.nhs.uk/catalogue/PUB20200/pres-cost-anal-eng-2015-apx.pdf (accessed 28 Sep 2017).

12 Curtis HJ. Prescription Cost Analysis 1998-2016 - drugs matched to BNF via fuzzy lookup. *figshare* Published Online First: 28 September 2017. doi:10.6084/m9.figshare.5450323.v1

13 Tidy C. Name Changes of Medicines. Patient Platform Limited. 2015.https://patient.info/health/name-changes-of-medicines (accessed 20 Sep 2017).

14 Curtis HJ. Prescription Cost Analysis 1998-2016 data processing and normalisation. 2017.https://gist.github.com/HelenCEBM/192307b3c671a391f5ad6b44a3676880 (accessed 28 Sep 2017).

15 Office for National Statistics. Population Estimates. ONS Statistical bulletin. 2017.https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2016 (accessed 11 Sep 2017).

16 Office for National Statistics. CPI All Items Index. ONS Time Series Data. 2017.https://www.ons.gov.uk/economy/inflationandpriceindices/timeseries/d7bt/mm23 (accessed 20 Sep 2017).

17 RStudio-Inc. Shiny. 2017.https://shiny.rstudio.com/ (accessed 27 Sep 2017).

18 Bostock M. D3.js - Data-Driven Documents. 2017.https://d3js.org/ (accessed 27 Sep 2017).

19 Curtis HJ, Goldacre B. NHS prescription cost analysis data 1998-2016. *figshare* Published Online First: 28 September 2017. doi:10.6084/m9.figshare.5447194.v1

20 Bateman DN, Good AM, Afshari R, *et al.* Effects of licence change on prescribing and poisons enquiries for antipsychotic agents in England and Scotland. *Br J Clin Pharmacol* 2003;**55**:596–603.

21 Szatkowski L, Coleman T, McNeill A, *et al.* The impact of the introduction of smoke-free legislation on prescribing of stop-smoking medications in England. *Addiction* 2011;**106**:1827–34.

22 Langley TE, Huang Y, McNeill A, *et al.* Prescribing of smoking cessation medication in England since the introduction of varenicline. *Addiction* 2011;**106**:1319–24.

23 Szatkowski L, Aveyard P. Provision of smoking cessation support in UK primary care: impact of the 2012 QOF revision. *Br J Gen Pract* 2016;**66**:e10–5.

24 Acheampong P, Cooper G, Khazaeli B, *et al.* Effects of MHRA drug safety advice on time trends in prescribing volume and indices of clinical toxicity for quinine. *Br J Clin Pharmacol* 2013;**76**:973–9.

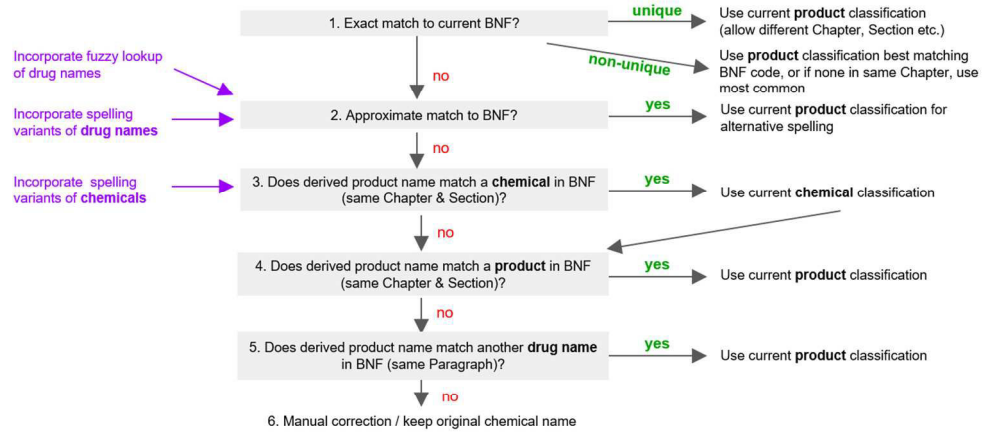
25 NHS-Digital. Prescriptions Dispensed in the Community, England 2006 to 2016. *Prescribing and Medicines Team, NHS Digital, UK* Published Online First: 29 June 2017.https://digital.nhs.uk/catalogue/PUB30014

LEGENDS TO FIGURES

Figure 1. *BNF code normalisation process flow chart describing how drug names were matched to the current BNF. Examples of matches at each numerated stage are given in table (bottom), with code/name changes in red. The first example demonstrates a drug which was matched to the current BNF through an exact match by name, but had moved from Chapter 4 to Chapter 2. For other types of matching, care was taken to avoid mistakenly matching to similar (but distinctly different) drugs across Chapters/Sections.*

Figure 2. *Screenshots from Trends tool, showing items per 1,000 population and inflation-corrected costs per 1,000 population for selected drugs. Product names are in parentheses. (a) Prescribing trends for four selected antipsychotic chemicals following the safety alert on thioridazine [20]. Full dashboard available at https://public.tableau.com/shared/XX7DTWSG2?:display_count=yes. (b) Prescribing trends for selected antidepressant chemicals [5]. Full dashboard available at https://public.tableau.com/shared/72SJGGP89?:display_count=yes. (c) Prescribing trends for all chemicals within the Paragraph of Nicotine Dependence (smoking cessation medications). Full dashboard available at https://public.tableau.com/shared/6BW9J5RJB?:display_count=yes.*

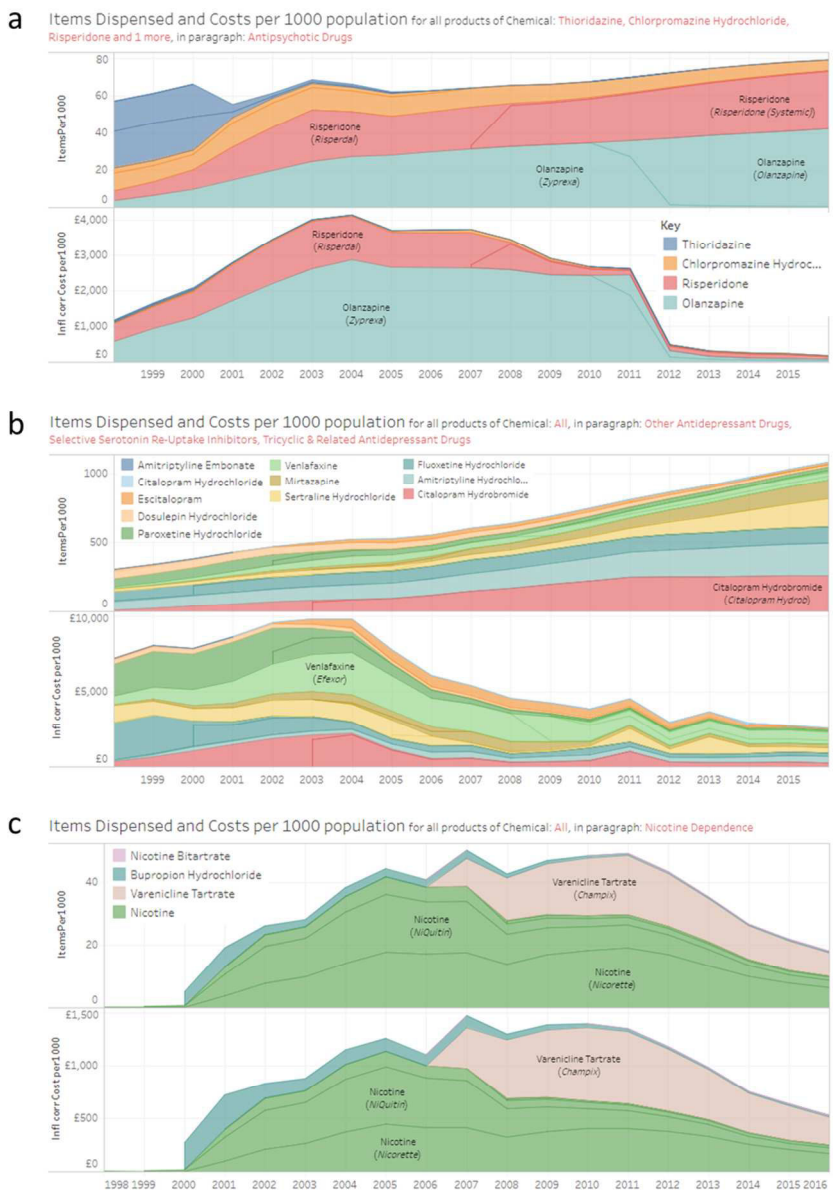
Figure 3. *Screenshots from Trends tool, showing inflation-corrected costs per 1,000 population (a) by Chapter, and (b) by Section for Chapter 2 (Cardiovascular System).*



Match Type	Drug Name Original	Chemical Original	BNF Code Original	Drug Name Current	Chemical Current	Code Derived	Product Current
1	Aspirin_Pdr Sach 37.5mg	Aspirin	0407010	Aspirin_Pdr Sach 37.5mg	Aspirin	0209000A0AA	Aspirin (Antiplatelet)
2	Frusemide_Tab 20mg	Furosemide	0202020	Furosemide_Tab 20mg	Furosemide	0202020L0AA	Furosemide
2	Amlodipine Besyl_Liq Spec 10mg/5ml	Amlodipine Besylate	0206020	Amlodipine_Oral Soln 5mg/5ml	Amlodipine	0206020A0AA	Amlodipine
3	Felendil XL_Tab 10mg P/R	Felodipine	0206020	(none)	Felodipine	0206020F0	Felendil XL
4	Cozaar-Comp_Tab	Losartan Potassium With Diuretic	0205052	(none)	Losartan Potassium With Diuretic	0205052P0BB	Cozaar-Comp
5	Atorvastatin_Liq Spec 20mg/5ml	Atorvastatin	0212000	Atorvastatin_Oral Soln 20mg/5ml	Atorvastatin	0212000B0AA	Atorvastatin
6	Terbut Sulph_Syr 1.5mg/5ml S/F	Terbutaline Sulphate	0301011	(none)	Terbutaline Sulfate	0301011V0AA	Terbut Sulf

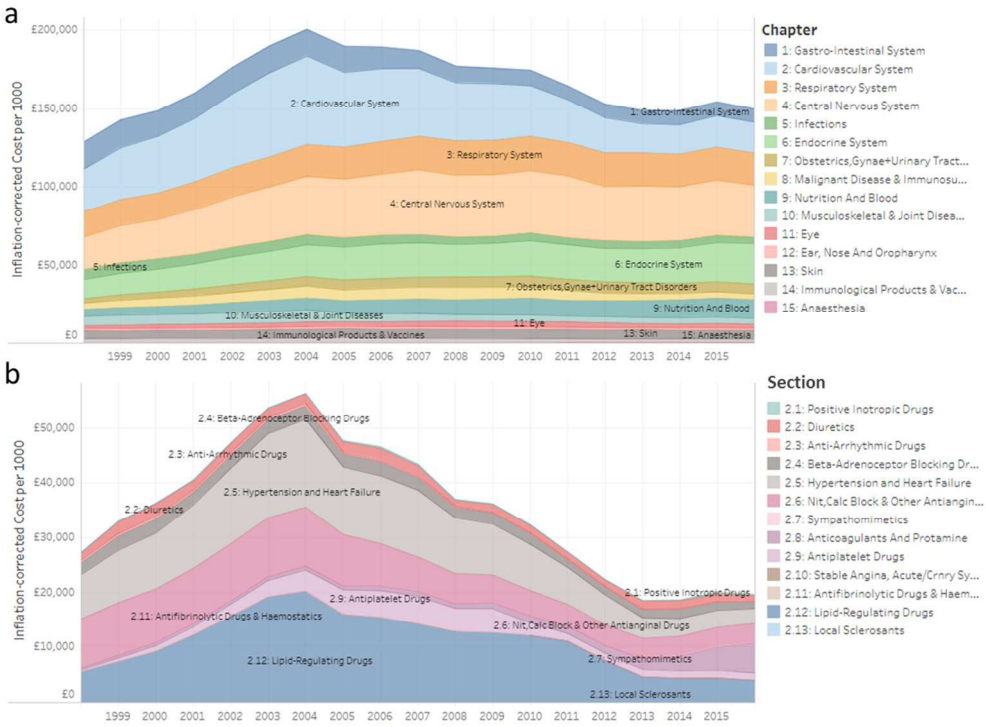
BNF code normalisation process flow chart describing how drug names were matched to the current BNF. Examples of matches at each numerated stage are given in table (bottom), with code/name changes in red. The first example demonstrates a drug which was matched to the current BNF through an exact match by name, but had moved from Chapter 4 to Chapter 2. For other types of matching, care was taken to avoid mistakenly matching to similar (but distinctly different) drugs across Chapters/Sections.

116x70mm (300 x 300 DPI)



Screenshots from Trends tool, showing items per 1,000 population and inflation-corrected costs per 1,000 population for selected drugs. Product names are in parentheses. (a) Prescribing trends for four selected antipsychotic chemicals following the safety alert on thioridazine [20]. Full dashboard available at https://public.tableau.com/shared/XX7DTWSG2?:display_count=yes. (b) Prescribing trends for selected antidepressant chemicals [5]. Full dashboard available at https://public.tableau.com/shared/72SJGGP89?:display_count=yes. (c) Prescribing trends for all chemicals within the Paragraph of Nicotine Dependence (smoking cessation medications). Full dashboard available at https://public.tableau.com/shared/6BW9J5RJB?:display_count=yes.

127x179mm (300 x 300 DPI)



Screenshots from Trends tool, showing inflation-corrected costs per 1,000 population (a) by Chapter, and (b) by Section for Chapter 2 (Cardiovascular System).

127x91mm (300 x 300 DPI)

SUPPLEMENTARY FILE

OpenPrescribing: Normalised Data and Software Tool to Research Trends in English NHS Primary Care Prescribing 1998-2016

Contents:

Table S1. Summary of original (unprocessed) PCA data by year (drugs in Chapters 1-15 only).

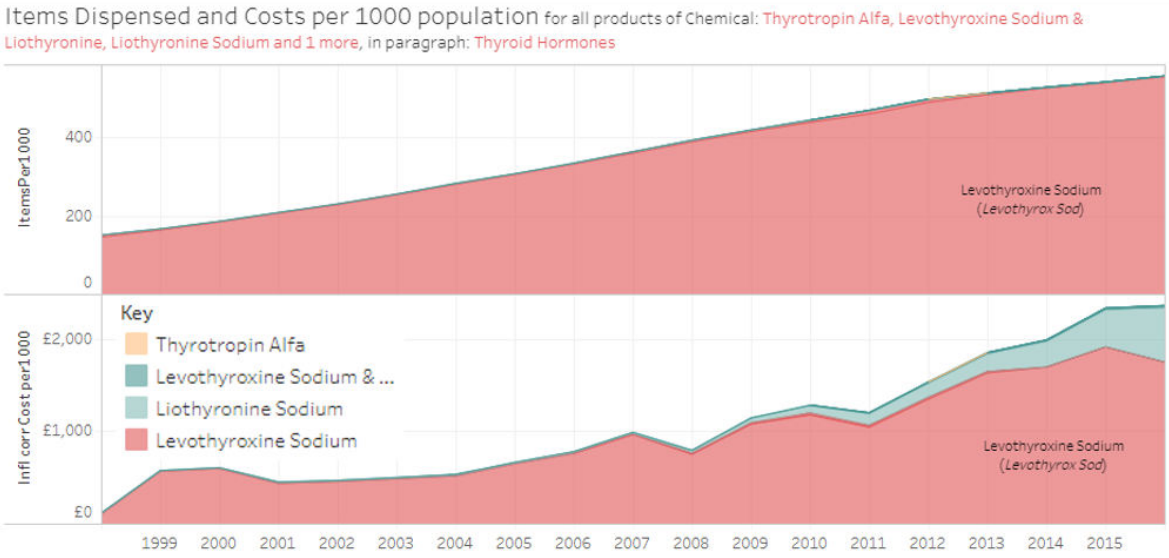
Year	Distinct count of Drug Name	Items	Cost
1998	6,338	497M	£4,440M
1999	6,587	513M	£5,011M
2000	6,613	535M	£5,284M
2001	6,754	569M	£5,784M
2002	6,834	599M	£6,487M
2003	6,893	630M	£7,113M
2004	6,912	666M	£7,645M
2005	6,907	699M	£7,452M
2006	6,810	728M	£7,660M
2007	7,056	772M	£7,810M
2008	7,202	817M	£7,716M
2009	7,401	859M	£7,892M
2010	11,703	898M	£8,162M
2011	11,751	932M	£8,101M
2012	12,207	969M	£7,802M
2013	12,318	996M	£7,846M
2014	12,576	1,027M	£8,022M
2015	12,875	1,043M	£8,403M
2016	13,285	1,061M	£8,284M
Total	***27,473***	**14,811M**	**£136,914M**

Table S2. Summary of drug code changes within the 1998-2016 prescribing datasets, by year.

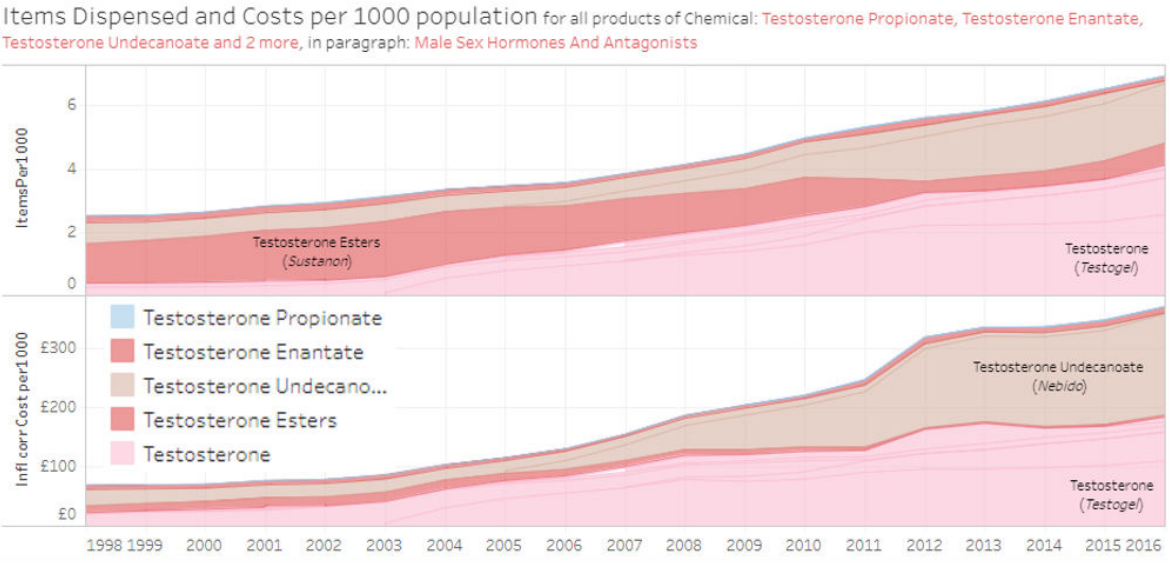
	Code change							
year	Chapter	Section	Paragraph	Sub-Paragraph	no change	no product match	no chemical match	Grand Total
1998	21	9	219	140	5,609	230	110	**6,338**
1999	21	7	188	88	5,917	237	129	**6,587**
2000	22	8	189	86	5,928	260	120	**6,613**
2001	26	9	178	85	6,069	278	111	**6,756**
2002	31	10	182	77	6,173	274	87	**6,834**
2003	34	9	191	79	6,256	241	83	**6,893**
2004	36	16	191	77	6,265	240	87	**6,912**
2005	26	14	157	85	6,315	245	65	**6,907**

2006	27	17	158	72	6,245	232	59	**6,810**
2007	35	12	168	74	6,483	237	47	**7,056**
2008	34	12	169	74	6,648	231	34	**7,202**
2009	3	4	180	76	6,919	219		**7,401**
2010	12	8	207	14	11,182	279	1	**11,703**
2011	15	4	130	14	11,331	257		**11,751**
2012	8	4	4	16	11,946	229		**12,207**
2013	2	4	8	1	12,139	163	1	**12,318**
2014		6	15	1	12,415	139		**12,576**
2015		6	6		12,770	93		**12,875**
2016		6	7		13,241	31		**13,285**

a


Items Dispensed and Costs per 1000 population for all products of Chemical: Thyrotropin Alfa, Levothyroxine Sodium & Liothyronine, Liothyronine Sodium and 1 more, in paragraph: Thyroid Hormones
ItemsPer1000
400
200
0
Levothyroxine Sodium
(Levothyrox Sod)
Infl corr Cost per1000
£2,000
£1,000
£0
Key
Thyrotropin Alfa
Levothyroxine Sodium & ...
Liothyronine Sodium
Levothyroxine Sodium
Levothyroxine Sodium
(Levothyrox Sod)
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

b


Items Dispensed and Costs per 1000 population for all products of Chemical: Testosterone Propionate, Testosterone Enantate, Testosterone Undecanoate and 2 more, in paragraph: Male Sex Hormones And Antagonists
ItemsPer1000
6
4
2
0
Testosterone Esters
(Sustanon)
Testosterone
(Testogel)
Infl corr Cost per1000
£300
£200
£100
£0
Testosterone Propionate
Testosterone Enantate
Testosterone Undecano...
Testosterone Esters
Testosterone
Testosterone Undecanoate
(Nebido)
Testosterone
(Testogel)
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

c

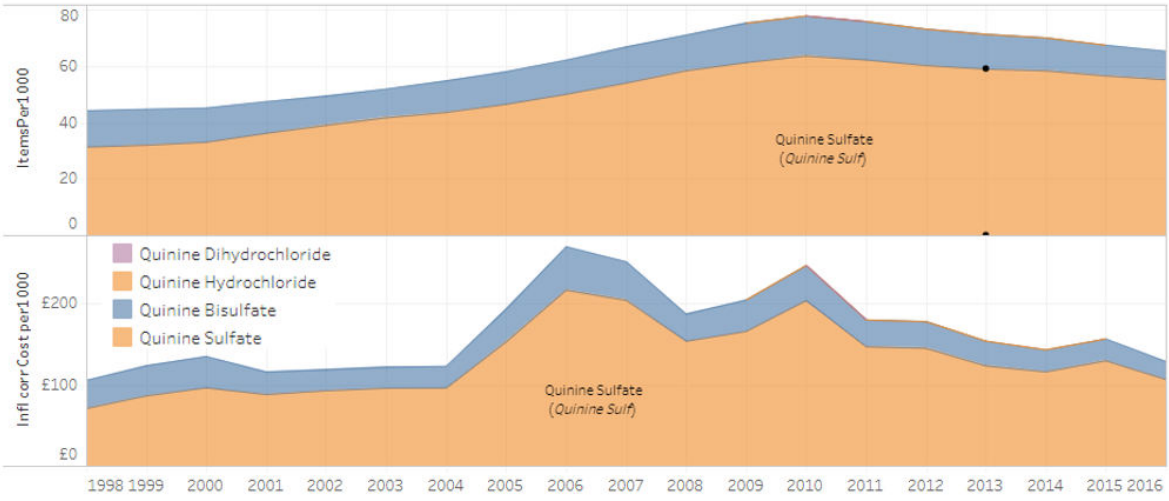

Items Dispensed and Costs per 1000 population for all products of Chemical: Quinine Dihydrochloride, Quinine Hydrochloride, Quinine Bisulfate and 1 more, in paragraph: Antimalarials
ItemsPer1000
80
60
40
20
0
Quinine Sulfate
(Quinine Sulf)
Infl corr Cost per1000
£200
£100
£0
Quinine Dihydrochloride
Quinine Hydrochloride
Quinine Bisulfate
Quinine Sulfate
Quinine Sulfate
(Quinine Sulf)
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

Figure S1. Screenshots from Trends tool, showing items per 1,000 population and inflation-corrected costs per 1,000 population for selected drugs. Product names are in parentheses. (a) Prescribing trends for all chemicals within the Paragraph of Thyroid Hormones. Full dashboard available at https://public.tableau.com/shared/GPW28PWJY?:display_count=yes. (b) Prescribing trends for all testosterone chemicals within the Paragraph of Male Sex Hormones. Full dashboard available at https://public.tableau.com/shared/YQ3ZFB3HY?:display_count=yes. c) Prescribing trends for all chemical forms of quinine (all of which are in the Antimalarials Paragraph). Full dashboard available at https://public.tableau.com/shared/85KJ2ZFN4?:display_count=yes.

Appendix - SQL Code for Processing and Normalisation of PCA data

A - Lookup Tables

A1. The special_cases lookup table

This is a workaround to assign a 'most likely' classification to the few problematic drug names which exist multiple times in BNF.

Lookup table is created by running the following script:

```
WITH temp as (
SELECT SUBSTR(SECTION_CODE,1,2) as chapter, section_code, presentation,
COUNT(DISTINCT product_code) as num
FROM ( SELECT DISTINCT section_code, section, para, subpara, chemical, product,
product_code, presentation FROM ebmdatalab.hscic.bnf )
GROUP BY chapter, section_code, presentation
HAVING num >1   --where name maps to more than one bnf code
ORDER BY chapter, num DESC)

SELECT
section_code, section, para, subpara, chemical, product, product_code,
presentation  -- this level is to filter to the top-prescribed code for each
drug name (according to latest detailed monthly data) (or, if none were
prescribed, then the first product name alphabetically
FROM (
  SELECT  -- this level joins all possible product codes to aggregated
prescribing data (2011-16) and ranks by items prescribed.
    a.*, b.items AS items_2011_2016,
    row_number() OVER (PARTITION BY a.presentation ORDER BY b.items DESC) AS
ranking -- We can use this to select the top/most likely drug code
  FROM (  -- this level is to look up all possible product codes for each drug
name in current BNF
      SELECT
```

```
      DISTINCT -- here we just want to go to product level rather than
individual presentations
      presentation,
      chapter, chapter_code,
      section, section_code,
      para, para_code,
      subpara, subpara_code,
      chemical, chemical_code,
      product, product_code
      from ebmdatalab.hscic.bnf where presentation in (select presentation from
temp where chapter < '18')

      ) a  -- now join to aggregated dataset grouped up to product level:
      LEFT JOIN ( SELECT substr(bnf_code,1,11) AS product_code, sum(items) as
items from ebmdatalab.aggregated_data.all_prescribed_BNFs_UpToSept2016 GROUP BY
product_code ) b
            ON a.product_code = substr(b.product_code,1,11)
      )
WHERE ranking = 1
ORDER BY chapter_code, presentation, product_code
```

A2. The lookup table of alternative drug spellings found within the data is created using the script below and saving as **ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_HC**

```
-- find drug name changes in PCA data to 2016
-- save results as ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_HC

WITH
a AS  (
  SELECT
      IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4)) AS
section_code, -- extra clause added to deal with those with extra spaces
2017-08-01

SUBSTR(drug_name,1,IF(STRPOS(drug_name,'_')>0,STRPOS(drug_name,'_')-1,length(dr
ug_name))) AS drug_name_part,  --take first part of drug name, up to underscore
(if there is one)
      MIN(year) AS min_year,  --this will help us to see which are the older vs
newer spellings used
      MAX(Year) AS max_year,
      SUM(items) AS Items
  FROM ebmdatalab.hscic.prescribing_pca_1998_2016_full
  WHERE  IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,2),SUBSTR(bnf_7_char,1,2))
< '18'
  GROUP BY
```

```
  section_code,
  drug_name_part
  ),

 b AS
(SELECT DISTINCT
section_code,
drug_name_part,
REPLACE(REPLACE(drug_name_part, 'i', '__'),'y','__') AS IY,
REPLACE(REPLACE(drug_name_part, 's', '__'),'z','__') AS SZ,
REPLACE(REPLACE(drug_name_part, 'ph', '__'),'f','__') AS PHF,
CONCAT(drug_name_part,'e') AS E,   -- add and E on to the end (note this only
works for the LAST word)
REPLACE(drug_name_part, ' ', ' e') AS E_mid -- add an E on to the end of all
words occurring before a space
FROM a)


SELECT
a.section_code,
a.drug_name_part,
CAST(a.min_year AS STRING) AS start_date,
CAST(a.max_year AS STRING) AS end_date,
a.items,
b.drug_name_part AS alternative,
CASE WHEN REPLACE(REPLACE(a.drug_name_part, 'i', '__'),'y','__') = b.IY THEN
'i_y'
      WHEN  REPLACE(REPLACE(a.drug_name_part, 's', '__'),'z','__') = b.SZ THEN
's_z'
      WHEN  REPLACE(REPLACE(a.drug_name_part, 'ph', '__'),'f','__') = b.PHF
THEN 'ph_f'
      WHEN  a.drug_name_part = b.E OR CONCAT(a.drug_name_part,'e') =
b.drug_name_part THEN 'e_end'
      WHEN  a.drug_name_part = b.E_mid OR REPLACE(a.drug_name_part, ' ', 'e ')
=b.drug_name_part THEN 'e_end'
      END AS type
FROM a
INNER JOIN b
  ON (REPLACE(REPLACE(a.drug_name_part, 'i', '__'),'y','__') = b.IY
      OR REPLACE(REPLACE(a.drug_name_part, 's', '__'),'z','__') = b.SZ
      OR REPLACE(REPLACE(a.drug_name_part, 'ph', '__'),'f','__') = b.PHF
      OR a.drug_name_part = b.E   --note this will only show this match once,
so we put in the other way around also
      OR a.drug_name_part = b.E_mid
      OR CONCAT(a.drug_name_part,'e') = b.drug_name_part
      OR REPLACE(a.drug_name_part, ' ', 'e ') =b.drug_name_part)
  AND a.drug_name_part != b.drug_name_part
  AND a.section_code = b.section_code
```

```
ORDER BY items desc
```

A3. The lookup table of Chemical name changes is created using the script below and saving as **pca_chemical_old_to_new_lookup**

```
-- PCA data - finding up to date chemical to combine with dataset
--save results as ebmdatalab.hscic.pca_chemical_old_to_new_lookup
WITH A as (
  SELECT
    IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4)) AS
Section,
    drug_name,
    count(distinct chemical) AS chems,
    max(year) AS Max_year_overall
    FROM
    ebmdatalab.hscic.prescribing_pca_1998_2016_full
    where
IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,2),SUBSTR(bnf_7_char,1,2)) <'18'
    GROUP BY
    Section,
    drug_name
    HAVING chems >1
    ),

B AS (
SELECT IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4))
AS section, drug_name, chemical,
    min(year) AS Min_year,
    max(year) AS Max_year
    FROM ebmdatalab.hscic.prescribing_pca_1998_2016_full
GROUP BY Section, drug_name, chemical
    ),

C AS (
SELECT DISTINCT
    A.Section,
    A.drug_name,
    B.chemical,
    B.min_year,
    B.max_year,
    IF(max_year = Max_year_overall,1,0) AS latest
    FROM A LEFT JOIN B ON A.drug_name = B.drug_name AND A.Section = B.Section
ORDER BY drug_name, chemical
)

SELECT old.section, old.drug_name, old.chemical AS old_chemical_name,
    nw.chemical AS new_chemical_name, nw.min_year AS Since
    FROM c old
```

```
     LEFT JOIN c nw ON old.drug_name = nw.drug_name AND old.chemical !=
nw.chemical and nw.latest = 1
    WHERE old.latest = 0
ORDER BY old.section, old.drug_name
```

A4. Known drug name changes

As reported online by patient.info
`ebmdatalab.hscic.drug_name_changes_2013`

A5. Fuzzy lookup for drugs not matching to BNF

List of drugs not matching BNF, identified through earlier iterations of the code.
These 1,084 drugs were matched to similar BNF names via fuzzy lookup in Excel and manually checked by a pharmacist.

List available at:
https://docs.google.com/spreadsheets/d/1UweKIZOLrKEzCtLULk5R5kJ4UyFIttvEouQ7RFGYrEA/edit#gid=594622641
and stored as `ebmdatalab.hscic.pca_bnf_name_to_code_fuzzy_lookup`

B - Data Extraction And Normalisation

B1. The latest chemical name for each drug is appended into the full dataset, to create prescribing_pca_1998_2016_full_v2

This does not take into account spelling changes but those will be handled later

```
-- save results as ebmdatalab.hscic.prescribing_pca_1998_2016_full_v2

SELECT a.*, COALESCE(c2.new_chemical_name, a.chemical) AS new_chemical_name
FROM
ebmdatalab.hscic.prescribing_pca_1998_2016_full a

LEFT JOIN ebmdatalab.hscic.pca_chemical_old_to_new_lookup c2
   ON a.drug_name = c2.drug_name
   AND a.chemical = c2.old_chemical_name
   AND IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,4),SUBSTR(bnf_7_char,1,4)) =
c2.section
```

C2. Run final data extraction parts 1 and 2 (scripts copied and updated from Issues #6 and #7)

B2a. Part 1

-- Final PCA data extraction part 1 (2016)

```
-- save results as ebmdatalab.tmp_eu.trends_from_pca

WITH
temp AS
(SELECT DISTINCT X.section_code, X.drug_name_part AS old_name, X.alternative
FROM
ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_HC X
INNER JOIN ebmdatalab.hscic.drug_name_alt_spellings_in_PCA_data_2016_HC Y ON
X.alternative = Y.drug_name_part AND Y.end_date = '2016'
),

b AS (
  SELECT DISTINCT
   chapter_code, chapter, section_code, section, para, subpara, chemical,
product, product_code,
   REPLACE(presentation,'GlucOsamine','prop-GlucOsamine') AS
presentation,REPLACE(presentation,' ','')
     AS presentation_no_spaces
   FROM ebmdatalab.hscic.bnf
   WHERE presentation NOT IN (SELECT presentation from
ebmdatalab.hscic.bnf_name_to_product_special_cases_helen)
   AND chapter_code <'18'),

a0 AS (
SELECT *,
      TRIM(bnf_7_char) AS bnf_7_char_trim,

SUBSTR(drug_name,1,IF(STRPOS(drug_name,'_')>0,STRPOS(drug_name,'_')-1,length(dr
ug_name)))
        AS drug_name_part,
      SUBSTR(drug_name,1,IF(STRPOS(drug_name,' ')>0,STRPOS(drug_name,'
')-1,length(drug_name)))
        AS drug_name_part_short,
      SUBSTR(chemical,1,IF(STRPOS(chemical,' ')>0,STRPOS(chemical,'
')-1,length(chemical)))
        AS chemical_short,
      REPLACE(drug_name,'GlucOsamine','prop-GlucOsamine') AS drug_name_a,

REPLACE(REPLACE(drug_name,'GlucOsamine','prop-GlucOsamine'),'Sulph','Sulf') AS
drug_name_b,
      REPLACE(new_chemical_name,'Sulph','Sulf') AS new_chemical_name_b
```

```
      FROM ebmdatalab.hscic.prescribing_pca_1998_2016_full_v2 a
      WHERE
IF(LENGTH(bnf_7_char)=9,SUBSTR(bnf_7_char,2,2),SUBSTR(bnf_7_char,1,2)) < '18'),

a1 AS (SELECT a0.*,
        z.new_bnf_code AS code_fuzzy,
        z.new_name AS drug_name_fuzzy,
        CONCAT( UPPER(substr(d.new_name,1,1)),
substr(D.new_name,2,LENGTH(D.new_name)-1) ) AS product_2013,
        E.alternative AS product_new_spelling,
        CONCAT( UPPER(substr(d1.new_name,1,1)),
substr(D1.new_name,2,LENGTH(D1.new_name)-1) ) AS chemical_2013,  -- note, this
capitalises the first letter only
        replace(a0.new_chemical_name_b,a0.chemical_short,D3.new_name) AS
chemical_2013b,
        replace(a0.drug_name_b,a0.drug_name_part,e.alternative) AS
converted_drug_name, -- incorporate new spellings into drug name
        replace(a0.drug_name_b,a0.drug_name_part,D.new_name) AS
converted_drug_name2,
        replace(a0.drug_name_b,a0.drug_name_part_short,D2.new_name) AS
converted_drug_name3,

SUBSTR(drug_name_b,1,IF(STRPOS(drug_name_b,'_')>0,STRPOS(drug_name_b,'_')-1,len
gth(drug_name_b))) AS drug_name_part_b

        FROM a0
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D ON
LOWER(drug_name_part) = D.old_name
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D1 ON
LOWER(a0.chemical) = D1.old_name
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D2 ON
LOWER(drug_name_part_short) = D2.old_name
      LEFT JOIN ebmdatalab.hscic.drug_name_changes_2013 D3 ON
LOWER(a0.chemical_short) = D3.old_name
      LEFT JOIN temp E ON a0.drug_name_part = E.old_name AND
SUBSTR(a0.bnf_7_char_trim,1,4) = e.section_code
      LEFT JOIN ebmdatalab.hscic.pca_bnf_name_to_code_fuzzy_lookup z ON
A0.drug_name = z.old_name
      ),


--CAPITALISE WHERE NEEDED:
A2 AS (
SELECT *,
      CONCAT( UPPER(substr(converted_drug_name2,1,1)),
substr(converted_drug_name2,2,LENGTH(converted_drug_name2)-1) ) AS
converted_drug_name4,
```

```
      CONCAT( UPPER(substr(converted_drug_name3,1,1)),
substr(converted_drug_name3,2,LENGTH(converted_drug_name3)-1) ) AS
converted_drug_name5,
      CONCAT( UPPER(substr(chemical_2013b,1,1)),
substr(chemical_2013b,2,LENGTH(chemical_2013b)-1) ) AS chemical_2013_c
--capitalise chemical as well.
      FROM A1
      ),

--COALESCE TO FORM "FINAL" NAMES
a3 AS (
 SELECT *,
COALESCE(converted_drug_name,converted_drug_name4,converted_drug_name5,drug_nam
e_fuzzy,drug_name_b)
   AS drug_name_F,
 COALESCE(product_new_spelling,product_2013,drug_name_part_b) AS
drug_name_part_F,
 COALESCE(chemical_2013, chemical_2013_c, new_chemical_name_b) AS chemical_F
 from A2
 ),

 --add a drug name field without spaces:
 a4 AS (
 select *, REPLACE(drug_name_F,' ','')  as drug_name_F_no_spaces
 from A3
 ),


a AS  (
  SELECT
      x.bnf_7_char_trim AS bnf_code,
      x.drug_name,
      drug_name_F,
      COALESCE(spc.presentation,b.presentation,ba.presentation,
bb.presentation,bc.presentation, bd.presentation)
        AS current_bnf_name,
      COALESCE(spc.product_code,b.product_code,ba.product_code,
bb.product_code,bc.product_code, bd.product_code)
        AS current_bnf_code,
      drug_name_part,
      drug_name_part_F, -- use as product name if no other
      x.section,
      x.subpara,
      x.chemical AS Chemical_original,
      x.chemical_F AS Chemical,
      x.Year,
      SUM(x.owc2) AS OWC2,   -- prescribed generically but no generic available
      SUM(x.NIC) AS Cost,
```

```
      SUM(x.items) AS Items,
      SUM(x.quantity) AS Quantity
  FROM a4 x
    --AND A.Currently_in_BNF = 'N'
    LEFT JOIN ebmdatalab.hscic.bnf_name_to_product_special_cases_helen spc ON
upper(x.drug_name_F) = upper(spc.presentation)  -- look up original drug
details in current bnf (drugs matching more than one drug in bnf)
    LEFT JOIN b ON upper(x.drug_name_F) = upper(b.presentation)   -- use upper
to match up examples like this: "Pentasa Sr_Tab 250mg" and "Pentasa SR_Tab
250mg"
            AND SUBSTR(x.bnf_7_char_trim,1,4) = b.section_code   -- look up
original drug details in current bnf.

    LEFT JOIN b ba ON upper(x.drug_name_F) = upper(ba.presentation)
            AND SUBSTR(x.bnf_7_char_trim,1,4) != ba.section_code  -- check if
drug now only belongs in a different section but same chapter
            AND SUBSTR(x.bnf_7_char_trim,1,2) = ba.chapter_code
    LEFT JOIN b bb ON upper(x.drug_name_F) = upper(bb.presentation)
            AND SUBSTR(x.bnf_7_char_trim,1,2) != bb.chapter_code  -- check if
drug now only belongs in a different chapter
    LEFT JOIN b bc ON upper(x.drug_name) = upper(bc.presentation) AND
b.presentation IS NULL AND ba.presentation IS NULL AND bb.presentation IS NULL
--also check original in case new drug name didn't work e.g. nifedipin(e)
            AND SUBSTR(x.bnf_7_char_trim,1,4) = bc.section_code
    LEFT JOIN b bd ON x.drug_name_F_no_spaces = bd.presentation_no_spaces   --
match without spaces e.g. Terbut Sulf_Inha 250mcg (400 D) vs "(400D)"
            AND SUBSTR(x.bnf_7_char_trim,1,4) = bd.section_code  AND
b.presentation IS NULL AND ba.presentation IS NULL AND bb.presentation IS NULL
-- look up original drug details in current bnf.

  GROUP BY
      bnf_code, drug_name, drug_name_F, current_bnf_name, current_bnf_code,
      drug_name_part, drug_name_part_F, -- use as product name if no other
      section, subpara, Chemical_original, Chemical, Year
      )

SELECT
    a.bnf_code,
    a.current_bnf_code AS Product_code_updated,
    SUBSTR(COALESCE(a.current_bnf_code,b.product_code,a.bnf_code),1,2) AS
Chapter_code_current,
    SUBSTR(a.bnf_code,1,2) AS BNF_Chap_Code,
    COALESCE(b.chapter, ch.description) AS Chapter_Current,
    ch.description AS Chapter_original,
    SUBSTR(COALESCE(a.current_bnf_code,a.bnf_code),3,2) AS
Section_code_current,
    SUBSTR(bnf_code,3,2) AS BNF_Section_Code,
    COALESCE(b.section,se.description,a.section) AS Section_Current,
```

```
    a.section AS Section_Original,
    SUBSTR(COALESCE(a.current_bnf_code,b.product_code,a.bnf_code),5,2) AS
Para_code_current,
    COALESCE(b.para,pa.description) As Para_current,
    COALESCE(b.subpara,a.subpara) As Subpara_current,
    a.subpara AS Subpara_original,
    COALESCE(b.chemical,a.chemical) As Chemical_current,
    a.Chemical_original,
    COALESCE(b.product, a.drug_name_part_F) AS Product_current,
    current_bnf_name,
    a.drug_name,
    IF(b.product_code IS NULL,'N','Y') AS Currently_in_BNF,
    a.year,
    a.Items,
    a.owc2,
    a.Quantity,
    a.Cost

  FROM a
  LEFT JOIN ebmdatalab.hscic.bnf_vertical ch ON SUBSTR(a.bnf_code,1,2) =
ch.code
  LEFT JOIN ebmdatalab.hscic.bnf_vertical se ON SUBSTR(a.bnf_code,1,4) =
se.code
  LEFT JOIN ebmdatalab.hscic.bnf_vertical pa ON SUBSTR(a.bnf_code,1,6) =
pa.code
  LEFT JOIN b ON a.current_bnf_name = b.presentation
            AND a.current_bnf_code = b.product_code
Save results as ebmdatalab.tmp_eu.trends_from_pca_2016
```

B2b. Part 2

```
-- final pca data extraction (2016) part 2
-- distinct product-chemical combinations in current BNF:
WITH
 chem_p AS (
   SELECT DISTINCT product, product_code, chemical_code, chemical,
   count (distinct product_code) Over (partition by chemical_code, product)
      AS Dist_prods_with_same_name
   FROM ebmdatalab.hscic.bnf
   WHERE chapter_code <'18'
   ORDER BY Dist_prods_with_same_name, product),

-- find all drug_name_parts in PCA which have been mapped to a new chemical:
chem_0 AS (
  SELECT
```

```
SUBSTR(drug_name,1,IF(STRPOS(drug_name,'_')>0,STRPOS(drug_name,'_')-1,length(dr
ug_name)))
      AS drug_name_part,
  drug_name, section, old_chemical_name, new_chemical_name
  FROM ebmdatalab.hscic.pca_chemical_old_to_new_lookup_2016),

-- distinct *chemicals* in current BNF:
chem_a AS (
SELECT chemical,
    count(distinct chapter) AS Chapters,
    count(distinct section) AS Sections,
    count(distinct para) AS Paras,
    count(distinct chemical_code) AS Codes,
    min(chemical_code) AS min_code
  FROM ebmdatalab.hscic.bnf
  WHERE chapter_code < '18'
  GROUP BY chemical
  ORDER BY codes DESC, paras DESC, chemical),

-- for chemicals with multiple codes:
-- check whether each chemical code is the only one in its paragraph / section
/ chapter
chem_a1 AS
(SELECT DISTINCT
  a.chemical, b.chemical_code, a.paras, b.para_code, a.sections,b.section_code,
a.chapters, b.chapter_code,
  count(distinct b.chemical_code) over (partition by b.chemical,chapter_code)
    AS appearances_by_chapter,
  count(distinct b.chemical_code) over (partition by b.chemical,section_code)
    AS appearances_by_section,
  count(distinct b.chemical_code) over (partition by b.chemical,para_code)
    AS appearances_by_para
  FROM ebmdatalab.hscic.bnf b
  INNER JOIN chem_a a ON a.chemical = b.chemical and a.codes > 1
  WHERE b.chapter_code < '18'
  ORDER BY chemical ),


-- SELECT ALL CHEMICALS FROM BNF WHICH MAP TO A SINGLE PRODUCT
-- used in final step only
b AS (
  SELECT DISTINCT
   chapter_code, chapter, section_code, section, para_code, para, subpara_code,
subpara, chemical_code
   FROM ebmdatalab.hscic.bnf
   WHERE chapter_code <'18'),
```

```
t as (
 SELECT t.*,
 REPLACE(c2.new_chemical_name,'Streptokinase-Streptodornase','Streptokinase &
Streptodornase')
      AS new_chemical_name,
 CASE WHEN Product_current LIKE 'Levonelle%' THEN '0703050A0BC'   --
'Levonelle'
      WHEN Product_current LIKE 'Postinor%' THEN '0703050A0BB'    -- 'Postinor'
      WHEN t.drug_name LIKE 'Terbut%Sulph_Syr%' THEN '0301011V0AA'  -- 'Terbut
Sulf'
      WHEN t.drug_name LIKE 'Thalidomide%' AND Chapter_code_current = '05' THEN
'0501100J0AA' -- 'Thalidomide (Antileprotic)'
      WHEN Product_current LIKE 'Menoring 50' THEN '0702010G0BE'  -- 'Menoring
50'
      WHEN t.drug_name = 'Acetylcy_Eye Dps 10% (Old)' THEN '1108010C0AA' --
'Acetylcy (Eye)'
      WHEN t.drug_name = 'Abilify Maintena_Inj 400mg Vl + Dil' THEN
'0402020ADBB' -- 'Abilify Maintena'
      WHEN Product_current LIKE 'Melatonin%' THEN '0401010ADAA' -- 'Melatonin'
      WHEN Product_current LIKE 'Varidase%' THEN '1311070R0BB' -- 'Varidase'
      WHEN t.drug_name = 'Cocois_Scalp Oint' THEN '1305020V0BB' -- 'Cocois'
      WHEN t.drug_name = 'Levocarnitine_Oral Soln Paed 1.5g/5ml30%' THEN
'0908010C0AA' -- 'Levocarnitine'
      ELSE product_code_updated
      END AS product_code_updated_manual,
 CASE WHEN Product_current LIKE 'Levonelle%' THEN 'Levonelle'
      WHEN Product_current LIKE 'Postinor%' THEN 'Postinor'
      WHEN t.drug_name LIKE 'Terbut%Sulph_Syr%' THEN 'Terbut Sulf'
      WHEN t.drug_name LIKE 'Thalidomide%' AND Chapter_code_current = '05' THEN
'Thalidomide (Antileprotic)'
      WHEN Product_current LIKE 'Menoring 50' THEN 'Menoring 50'
      WHEN t.drug_name = 'Acetylcy_Eye Dps 10% (Old)' THEN 'Acetylcy (Eye)'
      WHEN t.drug_name = 'Abilify Maintena_Inj 400mg Vl + Dil' THEN 'Abilify
Maintena'
      WHEN Product_current LIKE 'Melatonin%' THEN 'Melatonin'
      WHEN Product_current LIKE 'Varidase%' THEN 'Varidase'
      WHEN t.drug_name = 'Cocois_Scalp Oint' THEN 'Cocois'
      WHEN t.drug_name = 'Levocarnitine_Oral Soln Paed 1.5g/5ml30%' THEN
'Levocarnitine'
      ELSE product_current
      END AS product_current_manual
 FROM ebmdatalab.helen.trends_from_pca_2016 t
 LEFT JOIN chem_0 c2 ON t.drug_name = c2.drug_name AND t.chemical_current =
c2.old_chemical_name AND SUBSTR(t.bnf_code,1,4) = c2.section
 ),
```

```
  A AS (
 SELECT T.*,
 COALESCE(chem_p.product,c2.product,product_current_manual) AS current_product,
    -- use this order in coalesce because we want to update/replace any
existing product names for which we now have a better one.
 COALESCE(product_code_updated_manual,chem_p.product_code,c2.product_code)
    AS current_product_code,

COALESCE(chem_p.chemical,c2.chemical,chem_a.chemical,chem_a1.chemical,c3.chemic
al)
    AS unique_chem,  -- chemicals currently in BNF (uniquely)

COALESCE(chem_p.chemical_code,c2.chemical_code,chem_a.min_code,chem_a1.chemical
_code,c3.min_code)
    AS unique_chem_code

 FROM  t
  -- link to BNF using whole Product name (note this will be drug_name_part)
--------------------
  -- chemical must match as well because product names are not always unique.
  LEFT JOIN chem_p ON t.product_current = chem_p.product
            AND t.Product_code_updated_manual IS NULL
            AND (UPPER(chem_p.chemical) = UPPER(Chemical_current)
                      OR  UPPER(chem_p.chemical) = UPPER(new_chemical_name))
            AND SUBSTR(chem_p.chemical_code,1,6) = SUBSTR(bnf_code,1,6)
                      -- some chemicals sit in multiple paras.
            AND chem_p.Dist_prods_with_same_name = 1

  -- try shortening Product names in BNF to match products in data (only if
whole name is not found) --
    LEFT JOIN chem_p c2 ON t.product_current =
SUBSTR(c2.product,1,length(t.product_current))
            AND t.Product_code_updated_manual IS NULL
            AND chem_p.product IS NULL
            AND UPPER(c2.chemical) IN (UPPER(Chemical_current),
UPPER(new_chemical_name))
            AND SUBSTR(c2.chemical_code,1,6) = SUBSTR(bnf_code,1,6)  --some
chems sit in multiple paras.
            AND chem_p.Dist_prods_with_same_name = 1

  -- link to BNF using "original" chemical name for chemicals which are unique
in BNF  ---------------
  LEFT JOIN chem_a ON UPPER(chem_a.chemical)  = UPPER(Chemical_current)
            AND chem_a.codes = 1 AND chem_p.chemical IS NULL
            AND t.Product_code_updated_manual IS NULL
  -- link to BNF using NEW chemical name for chemicals which are unique in BNF
--------------------
  LEFT JOIN chem_a c3 ON c3.chemical  = new_chemical_name
```

```
            AND c3.codes = 1 AND chem_a.chemical IS NULL
            AND chem_p.chemical IS NULL
            AND t.Product_code_updated_manual IS NULL
  -- link to BNF using NEW chemical name for chemicals which are NON-unique in
BNF  --------------
  -- provided that no chemical has been assigned in a previous join.
  -- first check same paragraph then section then chapter.
  LEFT JOIN chem_a1 ON chem_a1.chemical  = Chemical_current
            AND chem_p.chemical IS NULL
            AND chem_a.chemical IS NULL
            AND c3.chemical IS NULL
            AND t.Product_code_updated_manual IS NULL
            AND (
              (chem_a1.para_code = SUBSTR(bnf_code,1,6) AND
chem_a1.appearances_by_para = 1)
              OR (chem_a1.section_code = SUBSTR(bnf_code,1,4) AND
chem_a1.appearances_by_section = 1)
              OR (chem_a1.chapter_code = SUBSTR(bnf_code,1,2) AND
chem_a1.appearances_by_chapter = 1)
                    )
 ORDER BY drug_name,year ),

 u AS (
 select bnf_code,
 Chapter_code_current,   BNF_Chap_Code,     Chapter_Current, Chapter_original,
 Section_code_current,   BNF_Section_Code,  Section_Current,   Section_Original,
 Para_code_current,      Para_current,
 Subpara_current,  Subpara_original,
 COALESCE(unique_chem_code, SUBSTR(Product_code_updated,1,9),
SUBSTR(current_Product_code,1,9))
   AS chem_code_today, --chemical code
 Chemical_original,
 COALESCE(unique_chem,Chemical_current) AS chem_today,
 COALESCE(Product_code_updated, current_product_code)    AS prod_code_today,
 COALESCE(current_product,     Product_current) AS prod_today,
   -- note this is opposite way around to code because we want to replace the
previous name
   -- but there may not be a code.
 current_bnf_name, drug_name,
 Currently_in_BNF,
 year,  Items,  owc2,  Quantity,  Cost
 FROM a
 ORDER BY drug_name, year)


SELECT
    bnf_code,
    COALESCE(b.chapter_code,Chapter_code_current) AS Chapter_code_current,
```

```
    BNF_Chap_Code,
    COALESCE(b.chapter,Chapter_Current) AS Chapter_Current,
    Chapter_original,
    COALESCE(SUBSTR(b.section_code,3,2),Section_code_current) AS
Section_code_current,
    BNF_Section_Code,
    COALESCE(b.section,Section_Current) AS Section_current,
    Section_Original,
    COALESCE(SUBSTR(b.para_code,5,2),Para_code_current) AS Para_code_current,
    COALESCE(b.para,Para_current) AS Para_current,
    COALESCE(b.subpara,Subpara_current) AS Subpara_current,
    Subpara_original,
    chem_code_today AS Chemical_code_current,
    Chemical_original,
    chem_today AS Chemical_current,
    prod_code_today AS Prod_code_current,
    prod_today AS product_current,
    current_bnf_name, drug_name,
    Currently_in_BNF, u.year,  Items,  owc2,  Quantity,  Cost,
    -- add calculated fields:
    1000*items/pop.Population AS ItemsPer1000,
    1000*quantity/pop.Population AS QuantityPer1000,
    Inf.Multiplier_2016*cost AS Infl_corr_Cost,
    1000*Inf.Multiplier_2016*cost/pop.Population AS Infl_corr_Cost_per1000,
    IEEE_DIVIDE(Inf.Multiplier_2016*Cost, Items) AS Infl_corr_CostPerItem,
    1000*owc2/pop.Population AS Owc2Per1000

 FROM U
 LEFT JOIN b ON u.chem_code_today = b.chemical_code
 LEFT JOIN ebmdatalab.ONS.england_midyear_population pop ON u.Year = pop.Year
 LEFT JOIN ebmdatalab.ONS.inflation_cpi inf ON u.Year = inf.Year
 --WHERE LENGTH(chem_code_today) =8
```