BMJ Open

The Evolution of Primary Care Databases in UK: A Scientometric Analysis of Research Output

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-012785
Article Type:	Research
Date Submitted by the Author:	25-May-2016
Complete List of Authors:	Vezyridis, Paraskevas; University of Nottingham, Business School Timmons, Stephen; University of Nottingham, Business School
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Evidence based practice
Keywords:	scientometrics, PRIMARY CARE, electronic patient records, STATISTICS & RESEARCH METHODS

The Evolution of Primary Care Databases in UK: A Scientometric Analysis of Research Output

Paraskevas Vezyridis[1*], Stephen Timmons[1]

[1] Centre for Health Innovation, Leadership and Learning (CHILL), Nottingham University Business School, Jubilee Campus, Nottingham NG8 1BB, UK

* Corresponding author

E-mail: Paraskevas.Vezyridis@nottingham.ac.uk

Telephone: +44 115 846 6370

Fax: +44 115 846 6667

Abstract

Objective To identify publication and citation trends, most productive institutions and countries, top journals, most cited articles and authorship networks from articles that used and analysed data from primary care databases (CPRD, THIN, QResearch) of pseudonymised electronic health records in UK.

Methods Descriptive statistics and scientometric tools were used to analyse a SCOPUS dataset of 1891 articles. Open access software was used to extract networks from the dataset (Table2Net), visualise and analyse co-authorship networks of scholars and countries (Gephi) and, density maps (VOSviewer) of research topics co-occurrence and journal co-citation.

Results Research output increased overall at a yearly rate of 18.65%. While medicine is the main field of research, studies in more specialised areas include biochemistry and pharmacology. Researchers from UK, USA and Spanish institutions have published the most papers. Most of the journals that publish this type of research and most cited papers come from UK and USA. Authorship varied between 3-6 authors. Keyword analyses show that smoking, diabetes, cardiovascular diseases and mental illnesses, as well as medication that can treat such medical conditions, such as non-steroid anti-inflammatory agents, insulin and antidepressants constitute the main topics of research. Co-authorship network analyses show that lead scientists, directors or founders of these databases are, to various degrees, at the centre of clusters in this scientific community.

Conclusions There is a considerable increase of publications in primary care research from electronic health records. The UK has been well placed at the centre of an expanding global scientific community, facilitating international collaborations and bringing together international expertise in medicine, biochemical and pharmaceutical research.

Strengths and limitations of this study

- First study to perform a scientometric analysis of research output from primary care databases of electronic patient records.
- We analysed articles published from 1995 to 2015 in order to explore the historical breadth and growth of this type of research.
- The analysis is limited on articles and structured data retrieved from the Scopus database.
- Some latest articles and related citations might not have appeared in Scopus when the dataset was extracted.

Introduction

Big data (analytics) refers to the aggregation and interrogation of – high volume, high velocity, high variety – datasets so as to reveal new, non-obvious, information and patterns [1]. This field is advancing because of technological and scientific developments in information infrastructure and digitisation [2]. For governments, opening up the datasets states hold about their citizens is believed to have, through computational and algorithmic analyses, a disruptive and transformative effect on knowledge [3]. In the United Kingdom (UK), big (open) data has been at the forefront of research activity and policy-making. Termed as one of the *eight great technologies* [4], UK has embraced the big (open) data movement more than many other developed countries (e.g. US, Australia, France) [3]. One area of particular relevance to big data analytics is healthcare.

In UK, the National Health Service (NHS) is organised around primary care and unless there is an accident or emergency involved whenever citizens would like to access the NHS they have to go through their primary care physician, known in the UK as a General Practitioner (GP). From there, they can be referred to a specialist at a hospital if it is deemed necessary. Secondary care clinicians can then feed back information to GPs. Since the vast majority of the population (98%) is registered with a general practice, GPs act not only as the main gatekeepers for the NHS but also as important custodians of a longitudinal electronic health record (EHR) [5]. There are now many ongoing primary care databases of anonymised patient records in UK that can be used for healthcare research. These population-based databases contain data originating from routine general practice. Some newly established databases and research platforms of linked EHRs include ResearchOne [6] and CALIBER [7]. While there are more than 9,600 general practices in UK that could potentially contribute data to these databases [8], it is usually 6-10% of these practices that do so.

Such databases are usually used for cross sectional surveys, case control or cohort studies and for epidemiological, drug safety, clinical and healthcare utilisation research purposes. They rely heavily on individual general practices voluntarily contributing data via the propriety clinical systems they use to maintain these patient records. The records are usually anonymised or pseudo-anonymised at source by allocating a unique number at each patient to allow for the updates of the records as well as for their linkage to other datasets, such as national mortality, national cancer registration and hospital records as well as with socioeconomic, ethnicity and environmental datasets. Access to these datasets is usually granted after scientific and ethics review and can be tailored to customer requirements. In this study, we examined the research output of three such databases that are well established in the research community and have contributed to a substantial number of

scientific studies and publications. These are the Clinical Practice Research Datalink (CPRD) [9], The Health Improvement Network (THIN) [10] and QResearch [11].

CPRD (formerly known as the General Practice Research Database) is a not-for-profit research service funded by the NHS National Institute for Health Research (NIHR) and the Medicines and Healthcare products Regulatory Agency (MHRA). It is owned by the UK Department of Health and contains the records of 11 million patients (4.4 million active) from 674 general practices [5]. There is a service cost associated with the preparation of the requested data. Contrary to the other services described below, CPRD does not extract data from a particular propriety clinical system. Any general practice can contribute data after a data sharing agreement with the software supplier. Importantly, it is the only database accessible online [12]. THIN database contains the health records of 12 million patients from around 600 general practices that use the Vision clinical system by INPS. IMS Health can provide access to data, for example, via a yearly sub-license to an academic institution. Compared to the other two databases, it is the only that can provide access to for-profit companies. QResearch is a research service located at the University of Nottingham. Its database contains the health records of 18 million patients from 1000 general practices that use the EMIS clinical system. Only academics employed by a UK university can have (in site) access only to a *sample* of the dataset (maximum 100,000 patients) that is sufficient enough to answer a specific research question/hypothesis. As this research service is not-for-profit and entirely self-funded there is usually a fee to be paid to cover the cost of the single data extraction.

Research-wise, the strengths of these databases lie in their size, breadth, representativeness of the UK population, long-term follow-up and data quality [5]. They include good information of morbidity and lifestyle, prescribing, preventive care, current standards of care and inter-practice variation [13]. Since they are continually (and automatically) updated, they are ideal for researchers to discover and monitor healthcare trends as well as the effectiveness of new interventions and treatments with minimum cost. They are increasingly linked to secondary care and mortality datasets. In contrast, their weaknesses include data extracted from propriety clinical systems developed for patient management and not for healthcare research. There are issues of missing data (e.g. from healthy patients), variable definitions for diagnoses, incomplete secondary care data (e.g. hospital admissions) and, health data (e.g. treatment adherence, over the counter medication) and data about sub-populations (e.g. prisoners, homeless, refugees, travellers) not captured adequately [5]. Information governance and informed consent procedures around the data sharing of EHRs for research are still considered complex [14]. These databases also require considerable clinical and scientific expertise and technical capacity in data management to support research. Importantly, when selecting a particular data resource for an observational

study, researchers have to consider several other factors, such as population covered and geographical distribution, data capture and latency, linkage with other resources, privacy and security, quality and validation [15].

In any case, these databases are highly regarded within the research community since they have proved their capability in helping researchers reach definitive answers to various healthcare debates of high public interest, particularly where other types of research have produced contradictory evidence. For example, in 2004 researchers from UK and Canada proved beyond doubt that measles-mumps-rubella (MMR) vaccination is not associated with autism in children [16]. In contrast to expensive, time-consuming, labour intensive and unrepresentative (of the population) traditional randomised trials [17], large-scale and randomised observational (comparative) evaluations of treatments and medications are minimally obtrusive for clinicians and patients and can support faster turnaround times of pragmatic evidence for use in clinical practice [18]. The aim of this study was to perform a scientometric analysis of articles, published from 1995 to 2015, which have used data from at least one of these primary care databases. To the best of our knowledge, this is the first study of a systematic mapping of primary care databases research output.

Methods

In this study, the Elsevier's Scopus database (www.scopus.com) was selected as the source of structured data of articles. This database covers more scientific articles than other databases (e.g. Thomson Reuters Web of Science) and has the advantage of providing advance export functionality of structured data about articles including full citation information, abstracts, keywords and references. On October 30, 2015, we searched, using the *document search* functionality, for all articles containing the terms "General Practice Research Database (GPRD)" OR "Clinical Practice Research Datalink (CPRD)" OR "The Health Improvement Network (THIN)" OR "QResearch" in Article Title, Abstract and Keywords.

The results were then limited to *articles*, *articles in press*, *conference papers*, *reviews*, *book chapters* and *short surveys*. *Notes*, *letters*, *editorials* and *errata* were excluded from the analysis. From there, we compared the resulted records with the bibliographic lists maintained by these databases [19–21] so as to include articles that could not be retrieved using the above search queries. Data cleansing included the removal of duplicate records and records missing vital information for the analysis (e.g. article title, journal etc). In the end, the fields of *authors*, *year*, *source title*, *affiliations*, *author keywords* and *document type* were used for the analysis.

The final bibliography retrieved from Scopus was imported to Table 2 Net [22] to extract networks of authors and contributing countries. It was then imported to Gephi [23] where the ForceAtlas 2 algorithm [24] was used to visualise the structural proximities for the communities of authors and contributing countries. The VOSviewer (version 1.6.3) [25] software was used to visualise bibliometric networks and densities [26] of frequent terms and journals. All other statistical analyses were performed using Microsoft Excel. We used the Journal Citation Reports (JCR) Science Edition 2014 to extract impact factor values for the identified journal titles.

Results

A total of 1,891 papers from 1995 to 2015 were included in this bibliometric and scientometric analysis. The results are presented below in graphs and tables.

Publication and citation trend

The literature related to the 3 primary care databases in England gradually increased from 7 in 1995 to 171 in 2015 (Table 1). The vast majority of papers are published in English. In total, these papers have already been cited 73,929 times. There is, however, a small percentage of 1.16% (n=163) papers that have not yet been cited yet. The average citation per year is approximately 3.52.

Year	No. of Papers	%	No. of Citations
2015	171	9.04	255
2014	214	11.31	1188
2013	203	10.73	1934
2012	175	9.25	3050
2011	148	7.82	3900
2010	129	6.82	5232
2009	126	6.66	5341
2008	115	6.08	4757
2007	96	5.07	5232
2006	74	3.91	5943
2005	81	4.28	5839
2004	73	3.86	6411
2003	46	2.43	2998

2002	52	2.74	3832
2001	51	2.69	3770
2000	51	2.69	6334
1999	26	1.37	1594
1998	29	1.53	2582
1997	17	0.89	1560
1996	7	0.37	1137
1995	7	0.37	1040
Total	**1891**	**100**	**73,929**

Table 1. Distribution of scientific literature by year

We explored the distribution of publications by document type is presented in Table 2 to identify the preferences of scholars using these databases in their research to share knowledge. The vast majority of scholars prefer to publish the findings of their research through journals, particularly as original articles (96.5%).

Type	No. of Papers	%
Article	1825	96.5
Conference Paper	18	0.95
Book Chapter	4	0.21
Review	41	2.16
Short Survey	3	0.15

Table 2. Distribution of scientific literature by Document Type

Next, we analysed the distribution of papers based on the academic discipline they have been categorised by Scopus (Table 3) and by which each paper may be attributed to more than one subject area (Scopus n.d.). Since we analyse bibliographic data based on published research using primary care databases, it comes as no surprise that the vast majority of papers are under the *Medicine* category. There is, however, a considerable number of papers (~25%) (also) under the categories *Biochemistry, Genetics and Molecular Biology* and *Pharmacology, Toxicology and Pharmaceutics*, which indicates a great emphasis on the use of these databases for the study of

medications and other substances in healthcare. It also indicates a potential interest of these databases from the Pharmaceutical sector.

Subject	No. of Papers	%
Medicine	1838	97.2
Biochemistry, Genetics and Molecular Biology	266	14.1
Pharmacology, Toxicology and Pharmaceutics	197	10.4
Neuroscience	78	4.1
Immunology and Microbiology	70	3.7
Agricultural and Biological Sciences	53	2.8
Nursing	44	2.3
Psychology	21	1.1
Arts and Humanities	13	0.7
Environmental Science	10	0.5

Table 3. Distribution of scientific literature by Discipline

Most Productive Institutions and Countries

For a deeper insight into contribution patterns and scientific impact, we first identified the top 10 institutions (by number of papers) authors have used as affiliation and then we analysed citation patterns (Table 4). We also analysed authors affiliations based on the country of their institution. The distribution of the top 10 contributing countries is presented in Table 5. Finally, we visualised the network of contributing countries using Gephi. We ended up with a network of 29 nodes and 175 edges (Fig 1). Each node represents a country, while its size denotes the country's degree and the colour the number of papers. The thickness of interconnected lines (edges) denotes the number of co-authored papers between the countries.

Rank	Institution	No. of Papers	%	Total Citations	Average Citations per Paper	Country
1	Boston University	217	11.15	12531	57.75	USA

2	University of Nottingham	209	10.73	7408	35.44	UK
3	Centro Espanol de Investigacion Farmacoepidemiologica (CEIFE)	151	7.76	7524	49.83	Spain
4	University College London	141	7.24	4135	29.33	UK
5	University of Utrecht	125	6.42	5390	43.12	Netherlands
6	London School of Hygiene & Tropical Medicine	119	6.11	3666	30.81	UK
7	University of Pennsylvania	110	5.65	7508	68.25	USA
8	King's College London	78	4.01	1722	22.08	UK
9	Medicines and Health Care Products Regulatory Agency	94	4.83	2623	27.90	UK
10	University of Oxford	86	4.42	1806	21.00	UK

Table 4. Most productive institutions

Rank	Country	No. of Papers	%
1	UK	1202	63.56
2	USA	563	29.77
3	Spain	192	10.15
4	Netherlands	164	8.67
5	Switzerland	115	6.08
6	Canada	112	5.92
7	Sweden	106	5.60
8	Germany	64	3.38
9	France	51	2.69
10	Italy	36	1.90

Table 5. Top contributing countries

[Fig 1. Network of contributing countries.]

The majority of the most productive institutions are universities. Top universities include Boston University, the University of Nottingham, UCL and the University of Utrecht. Apart from these academic institutions, a research unit in Spain (CEIFE) and an executive agency in UK (MHRA) are involved in primary care databases-based research. Switzerland, Canada, Sweden, Germany, France and Italy had no institution among the top ten list, although they were ranked among the top 10 productive countries. Most papers are published by scholars from UK (63.56%), followed by the United States (USA) and Spain. With the exception of USA and Canada, most of the productive countries are in Europe. What is particularly interesting in these two tables is that scholars in institutions from the USA and Spain produce not only a great number of publications but also publications that are widely recognised by this scientific community in terms of citations.

Taking into account the measurements of *Weighted Degree*, *Clustering*, *Eigen Vector Centrality* and *Betweeness* Centrality (Table 6) we observe that, once again the UK, followed by USA, is placed at the centre of this scientific community. With the highest degrees of all measurements, institutions from this country are the most well-connected and authoritative ones, facilitating the linking between institutions in other countries.

Rank	Country	Occurrences	Weighted Degree	Page Rank	Eigen Centrality	Closness Centrality	Betweeness Centrality
1	UK	5770	27.0	0.062	1.0	0.933	0.250
2	USA	2506	26.0	0.060	0.98	0.903	0.229
3	Netherlands	666	22.0	0.047	0.95	0.8	0.049
4	Canada	567	18.0	0.039	0.83	0.717	0.024
5	Switzerland	409	18.0	0.039	0.85	0.717	0.015
6	Italy	85	17.0	0.037	0.78	0.7	0.024
7	Spain	481	17.0	0.037	0.82	0.7	0.011
8	Australia	37	16.0	0.036	0.76	0.682	0.024
9	Sweden	315	16.0	0.036	0.77	0.682	0.023
10	Germany	121	16.0	0.036	0.71	0.682	0.020

Table 6. Top countries by centrality

Top Journals

In Table 7, we identify the top 10 journals where most research is published. Six of these journals are published by a UK publisher and the rest are published in the USA. The journal of *Pharmacoepidemiology and Drug Safety* features at the top of list, followed by the *British Journal of Clinical Pharmacology* and *Pharmacotherapy*, which signifies the focus of research, produced from these primary care databases on the safe use of medication. This focus can also be seen in Table 3, where (apart from medicine) most papers are published in the fields of *Biochemistry, Genetics and Molecular Biology* and *Pharmacology, Toxicology and*

Pharmaceutics (Scopus classification), but also in Table 5, where most of the top cited papers refer to potential risk for particular complications/conditions from the use of specific medication. More specialised journals, like *Diabetes Care* and *Annals of the Rheumatic Diseases*, are also featured in this list, indicating a particular focus of research activity in specific spectrums of diseases.

Rank	Journal Name	No. of Papers	%	Publisher	Impact Factor	Open Access	Country
1	Pharmacoepidemiology and Drug Safety	115	6.08	Wiley	2.939	Hybrid	UK
2	British Medical Journal	100	5.28	BMJ Publishing Group	17.445	Full	UK
3	British Journal of General Practice	67	3.54	Royal College of General Practitioners	2.294	Full	UK
4	British Journal of Clinical Pharmacology	57	3.01	Wiley	3.878	Hybrid	UK
5	Plos One	51	2.69	Public Library of Science	3.234	Full	USA
6	BMJ Open	41	2.16	BMJ Publishing Group	2.271	Full	UK
7	Pharmacotherapy	34	1.79	Wiley	2.662	Hybrid	USA
8	Diabetes Care	30	1.58	American Diabetes Association	8.420	Hybrid	USA
9	Epidemiology	24	1.26	Wolters Kluwer	6.196	Hybrid	USA
10	Annals of the Rheumatic Diseases	23	1.21	BMJ Publishing Group	10.377	Hybrid	UK

Table 7. Top journals of published research

Four journals in this list are open access (BMJ, British Journal of General Practice, Plos One, BMJ Open), which greatly facilitates the sharing of knowledge without limitations. The BMJ enjoys a widespread recognition of the high quality of its published studies as indicated by the high impact factor. The rest are behind a pay wall but

offer authors an open access option to publish their research (hybrid access). An extra column with the Impact Factors of these top 10 journals from the 2014 JCR was also added in the table.

What is also of particular importance in terms of scientific impact is that the 10 most cited papers identified below (see Table 8) have not been published in journals in this list. However, by performing a (full counting) analysis of co-citation links in VOSviewer (Fig 2) for journals cited in the Scopus dataset (minimum number of citations=10) we see that most of this list is represented here (blue - lowest density, red - highest density).

[Fig 2. Journal co-citation analysis.]

Most Cited Papers

Next, we focused on the top 10 papers [27–36] and calculated the total count of citations for each paper (Table 8) for the period 1995-2015. Citations totalled 7,194 (1.02%) of all citations in this dataset. It seems that these studies in dementia, psoriasis, fractures, cardiovascular diseases and gastrointestinal complications in relation to certain medications have been of great interest in this scientific community. The majority of the top 10 most cited papers (60%) are open access at the publisher's website and can be freely read by anyone.

Rank	Authors / Title	Year	Country	Journal	Impact Factor	Citations	Open Access
1	Jick H., Zornberg G.L., Jick S.S., Seshadri S., Drachman D.A. Statins and the risk of dementia	2000	USA	*Lancet*	45.217	1322	No
2	Gelfand J.M., Neimann A.L., Shin D.B., Wang X., Margolis D.J., Troxel A.B. Risk of myocardial infarction in patients with psoriasis	2006	USA	*Journal of the American Medical Association*	35.289	854	Yes
3	Van Staa T.P., Leufkens H.G.M., Abenhaim L., Zhang B., Cooper C. Use of oral corticosteroids and risk of fractures	2000	UK	*Journal of Bone and Mineral Research*	6.832	796	Yes

4	Henry D., Lim L.L.-Y., Rodriguez L.A.G., Perez Gutthann S., Carson J.L., Griffin M., Savage R., Logan R., Moride Y., Hawkey C., Hill S., Fries J.T. Variability in risk of gastrointestinal complications with individual non-steroidal anti-inflammatory drugs: Results of a collaborative meta-analysis	1996	Australia, Spain, USA, New Zealand, UK	*British Medical Journal*	17.445	688	Yes
5	Yang Y.-X., Lewis J.D., Epstein S., Metz D.C. Long-term proton pump inhibitor therapy and risk of hip fracture	2006	USA	*Journal of the American Medical Association*	35.289	637	Yes
6	Currie C.J., Poole C.D., Gale E.A.M. The influence of glucose-lowering therapies on cancer risk in type 2 diabetes	2009	UK	*Diabetologia*	6.671	629	Yes
7	Jick H., Jick S.S., Gurewich V., Myers M.W., Vasilakis C. Risk of idiopathic cardiovascular death and nonfatal venous thromboembolism in women using oral contraceptives with differing progestagen components	1995	USA, UK	*Lancet*	45.217	612	No
8	Dial S., Delaney J.A.C., Barkun A.N., Suissa S. Use of gastric acid-suppressive agents and the risk of community-acquired Clostridium difficile-associated disease	2005	Canada	*Journal of the American Medical Association*	35.289	582	Yes
9	Smeeth L., Thomas S.L., Hall A.J., Hubbard R., Farrington P., Vallance P. Risk of myocardial infarction and stroke after acute infection or	2004	UK	*New England Journal of Medicine*	55.873	546	Yes

	vaccination						
10	Neimann A.L., Shin D.B., Wang X., Margolis D.J., Troxel A.B., Gelfand J.M. Prevalence of cardiovascular risk factors in patients with psoriasis	2006	USA	*Journal of the American Academy of Dermatology*	4.449	528	No

Table 8. Most cited papers

Of the 10 papers, 8 are single country papers, while none were singled authored. Again, the USA has a considerable presence in this list, producing papers that are highly cited. In addition, many of the highly productive authors identified below (Table 10) were also found in this list.

Authorship Patterns and Networks

Authorship distributions varied from single to a maximum of 155 authors – for a study about the feasibility of international collaboration to evaluate, based on a common protocol, the risk of Guillain-Barré syndrome following pH1N1 vaccination [37]. In total, there were 9385 authors involved in the 1981 papers during 1995-2015. In Table 9, we can see that more than three-quarters of all papers were published by 3 or more authors. Only 1% of papers were written by a single author, while 5 papers didn't have any authorship details. Almost a quarter of all papers was published by 4 authors, which have been widely cited across this scientific community. This indicates the high degree of expert collaboration in this field that is perhaps necessary in analysing millions of primary care records.

Rank	No. of Authors	No. of Papers	%	Citations	%
1	4	460	24.33	16,437	22.23
2	5	383	20.25	16,071	21.74
3	3	310	16.39	12,209	16.51
4	6	252	13.33	12,084	16.35
5	2	145	7.67	7,159	9.68
6	7	123	6.50	3,798	5.14
7	8	88	4.65	2,914	3.94
8	>11	48	2.54	1,792	2.42
9	9	28	1.48	617	0.83
10	10	25	1.32	412	0.56
11	1	19	1.00	436	0.59

Table 9. Co-authorship distribution

Table 10 provides the ranking of the top 10 scholars, first, in terms of research productivity based on the overall number of co-authored papers. While, generally, most scholars are predominately from the UK, the Director of the Spanish Centre for Pharmacoepidemiologic Research (CEIFE) [38] is the scholar with the most published research out of these primary care databases. Also, there are researchers in this field who do not necessarily come from the academic environment. The pharmaceutical sector is actively involved in the knowledge production from electronic primary care records.

Rank	Author	Affiliation	Country	No. of Papers	Weighted Degree	Clustering	Eigen Centrality	Closness Centrality	Betweeness Centrality
1	Rodriguez, L.A.G.	Spanish Centre for Pharmacoepidemiologic Research (CEIFE)	Spain	166	82.0	0.082	0.346	0.337	0.073
2	Jick, S.S.	Boston University	USA	142	90.0	0.066	0.327	0.329	0.078
3	Van Staa, T.P.	University of Manchester	UK	115	144.0	0.063	1.0	0.403	0.217
4	Jick, H.	Boston University	USA	98	55.0	0.098	0.217	0.322	0.032
5	Meier, C.R.	University of Basel	Switzerland	94	54.0	0.116	0.232	0.294	0.014
6	Hubbard, R.	University of Nottingham	UK	86	82.0	0.092	0.441	0.359	0.067
7	Smeeth, L.	London School of Hygiene & Tropical Medicine	UK	79	89.0	0.101	0.578	0.366	0.067
8	Hippisley-Cox, J.	University of Nottingham	UK	72	57.0	0.144	0.244	0.325	0.065
9	Johansson, S.	AstraZeneca	Sweden	70	46.0	0.218	0.267	0.310	0.012
10	Cooper, C.	University of Southampton	UK	65	67.0	0.139	0.414	0.342	0.035
	West, J.	University of Nottingham	UK	65	42.0	0.213	0.230	0.310	0.012

Table 10. Most productive authors

Considering only those scholars who have co-authored at least 2 papers in this dataset, the analysis suggested a network (Fig 3) with 1261 nodes and 6186 edges. Here, each node represents an author, while its size denotes the number of author's papers. The interconnected lines (edges) denote the co-authored papers between the

authors. For better visualisation, we limited the number of minimum degrees to 5 (maximum degrees=145). After a modularity measurement, to identify community structure [39], we observe some established collaborative teams (clusters with different colours) around specific and highly productive scholars in the analysis of data from primary care databases also found in Table 10. We also observe a new (blue) cluster around the lead statistician for THIN [40] – one of the 3 primary care databases studied.

[Fig 3. Clustered co-author network.]

Taking into account the measurements of *Weighted Degree*, *Clustering*, *Eigen Vector Centrality* and *Betweeness Centrality* (Table 10), results indicate a cluster placed at the centre of this scientific community. With the lowest degree of clustering and the highest degrees of all the other measurements, its promiment scholar is the most well-connected, facilitating, more than any other scholar, linking between other scientific clusters and scholars.

What is also particularly interesting is the fact that that some of the scholars (and their institutions) in this list are affiliated, to a certain extent, to these databases, having served or currently acting as their founders, directors, lead scientists or members of their scientific committee [41,42]. For example, lead scientists from the Boston Collaborative Drug Surveillance Program in Boston University were among the first who developed the technical and scientific capacity of these databases in pharmacoepidemiological research [43,44].

Research Topics

We conducted a keyword analysis to identify important topics of published research. For this, we first extracted from the bibliographic dataset 5,813 unique keywords as indexed by Scopus [45] to base our analysis on more complete indexing information compared to authors' keywords. We retrieved the top 30 keywords for two specific categories: *medical conditions* and *medications/substances* (Tables 11 and 12). Next, we created a (full counting of) term co-occurrence density map in VOSviewer (Fig 4) by building a text corpus out of the title and abstract fields in the bibliographic dataset (minimum number of a term occurrence=10). In this way, we were able to identify topics that not only appear more frequently in the literature but that they are also strongly related to each other, forming certain clusters of topics. Blue colour indicates a low density of terms and red colour the highest density of terms. In many cases, the density map represents the frequency of indexed keywords in Tables 11 and 12. Clearly, smoking, diabetes, cardiovascular diseases, mental illnesses, psoriasis, obesity, pregnancy and cancer as well as medication and substances that can treat these medical conditions, such as aspirin, insulin,

antidepressants and non-steroid anti-inflammatory agents (NSAID), have been of great interest for scholars using EHRs in primary care.

Rank	Keyword	Occurrences	Rank	Keyword	Occurrences
1	smoking	328	16	cardiovascular diseases	96
2	diabetes mellitus	223	17	myocardial infarction	94
3	hypertension	223	18	chronic obstructive lung disease	90
4	non insulin dependent diabetes mellitus	179	19	heart failure	86
5	depression	167	20	cerebrovascular accident	85
6	stroke	165	21	rheumatoid arthritis	83
7	asthma	158	22	epilepsy	81
8	diabetes mellitus, type 2	155	23	breast cancer	78
9	cancer risk	150	24	fracture	75
10	cardiovascular risk	147	25	psoriasis	75
11	cardiovascular disease	133	26	gastrointestinal hemorrhage	69
12	obesity	129	27	hip fracture	68
13	heart infarction	126	28	osteoporosis	68
14	pregnancy	125	29	colorectal cancer	65
15	ischemic heart disease	104	30	fractures, bone	65

Table 11. Top keywords: medical conditions

Rank	Keyword	Occurrences	Rank	Keyword	Occurrences
1	nonsteroid antiinflammatory agent	182	16	proton pump inhibitor	75
2	acetylsalicylic acid	154	17	warfarin	71
3	metformin	150	18	antidiabetic agent	69

4	corticosteroid	143	19	anticonvulsive agent	68
5	hydroxymethylglutaryl coenzyme a reductase inhibitor	138	20	serotonin uptake inhibitor	65
6	insulin	133	21	calcium channel blocking agent	64
7	antidepressant agent	124	22	anti-bacterial agents	62
8	beta adrenergic receptor blocking agent	108	23	hydroxymethylglutaryl-coa reductase inhibitors	62
9	hypoglycemic agents	91	24	oral antidiabetic agent	61
10	anti-inflammatory agents, non-steroidal	90	25	paracetamol	56
11	dipeptidyl carboxypeptidase inhibitor	88	26	diuretic agent	53
12	antihypertensive agent	82	27	ibuprofen	52
13	neuroleptic agent	80	28	simvastatin	52
14	antibiotic agent	77	29	tricyclic antidepressant agent	52
15	hemoglobin a1c	75	30	diclofenac	50

Table 12. Top keywords: medications/substances

[Fig 4. Term co-occurrence density map.]

Discussion

This study identified the leading institutions, countries, authors, journals and topics as well as their networks of published research that have used primary care databases in the UK to extract and analyse data from EHRs. There is a growing production of such papers which indicates the interest of a global and highly collaborative scientific community in this field and also the knowledge and insights that can be gained for healthcare improvement. Publication output increased from 7 papers in 1995 to 171 by October 2015. We estimated the Compound Annual Growth Rate (CAGR), for the years 1995-2014, to be 18.65%. The vast majority of publications (96.5%) were journal articles. While this research field can be located, generally, in medicine, biochemical and pharmaceutical developments seem to be equally important so as to address specific and

widespread medical conditions, such as diabetes, cardiovascular diseases, mental illnesses, smoking, obesity and cancer.

The UK has been well placed in this scientific field. This is partly due to the fact that there is now more than 30 years of classified data available in GP information systems [46]. The investment in developing primary care databases from EHRs for research purposes has placed the country at the centre of an emergent network of collaborations across the globe, bringing together international expertise for the analysis of ever-expanding and increasingly interlinked clinical datasets from primary, secondary and tertiary care. Access to these datasets has also allowed researchers and institutions from other countries to develop their own programmes of research to answer important clinical questions. Six of the most productive institutions are located in the UK and 63.56% of publications were authored by scholars affiliated with this country, followed by the USA. Interestingly, the top institutions were not exclusively universities. Among the top list, we can find a research unit in Spain (CEIFE) and the executive agency in UK that funds and runs the CPRD database (MHRA), while one of the most productive scholars is affiliated with a pharmaceutical company. This signifies the great interest of various actors, from academic, governmental and private sectors, in research with primary care databases.

The geographical trend can also be observed from the location of journals with the most published papers. Six of the top 10 journals are published house in the UK, followed by the USA. The journals with the most papers published included *Pharmacoepidemiology and Drug Safety*, *BMJ*, *British Journal of General Practice* and *British Journal of Clinical Pharmacology*, which signifies the great interest of scholars in using data from EHRs for pharmaceutical research. This is partially because one of the oldest sets of routine information collected by GP practices in the UK and made available by these databases is drug histories [43]. Regarding restrictions on access to research outputs, only 4 journals in this list are fully open access. This may limit access to knowledge to researchers and members of the public that cannot afford subscription costs. Interestingly, it is the more established journals in medicine, like *JAMA*, *Lancet* and *NEJM*, that have published some of the most cited papers in this bibliometric dataset and enjoy a high level of co-citation activity.

Keyword analyses show that smoking, diabetes, cardiovascular diseases, mental illnesses, psoriasis, obesity, pregnancy and cancer constitute the main topics of research activity using EHRs in primary care. Often, this research concentrates on developing algorithms to identify risk of occurrence of a particular disease. Researchers are also interested in investigating certain medication and substances that can treat the medical conditions, such as aspirin and other non-steroid anti-inflammatory agents (NSAID), insulin and antidepressants.

For the vast majority of publications, authorship varied between 3-6 authors, indicating widely collaborative, international, efforts to promote research in this field. Co-authorship network analyses showed that lead scientists, directors and founders of these databases were found, to various degrees, at the centre of clusters in this scientific community, highlighting their invaluable contribution to knowledge production. As Azoulay et al. [47] have demonstrated in their study about eminent researchers and the vitality of a field, the development of co-authorship networks and clusters of collaborators in newly established scientific domains might be useful to boost research productivity. Based on each database's data access requirements, their established researchers appear to have a fundamental role in facilitating and promoting international collaborations for more researchers, institutions and countries to gain from the insights and outcomes of research with EHRs. Importantly, they have a clear and in-depth understanding and knowledge of the kind of research activities these databases can support in terms of data quality, structure and EHR coding practices. As these databases are expected to open up in the future to more stakeholders from various disciplines around health and as universities prepare to incorporate training in data science skills (e.g. statistics, biomedical informatics, biology and medicine) [48] into their clinical curricula, so as to nurture the next generation of clinical investigators [49], these established researchers could promote quality, reliable and ethically appropriate scientific research [50] from complicated and highly contextual datasets.

Our study has the typical limitations of a scientometric study. We analysed articles published in a period of 20 years in order to explore the historical breadth and growth of research from electronic primary care records. However, this analysis is limited on structured data retrieved from one bibliometric database of peer-reviewed literature. Therefore, only articles published in journals in its index were analysed. Also, some latest articles and related citations might not have been retrieved at the time of the search, which might explain the decrease in the number of publications and citations particularly from 2010 onwards.

In conclusion, output of primary care research from electronic health records has consistently increased since their development. The development of these databases in the UK has placed the country at the centre of an expanding global scientific community, facilitating international collaborations and bringing together international expertise in medicine, biochemical and pharmaceutical research.

Contribution Statement

PV and ST conceived and designed the study. PV collected and analysed the data. PV and ST interpreted the findings, drafted the manuscript and approved the final manuscript for submission.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by a Marie Sklodowska-Curie Individual Fellowship from European Commission (EC) (2014-IF-659478).

Data Sharing

No additional data available.

References

1 Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: ethical issues. London, UK: 2015.

2 Goffey A, Pettinger L, Speed E. Politics, Policy and Privatisation in the Everyday Experience of Big Data in the NHS. In: Hand M, Hillyard S, eds. *Big Data? Qualitative Approaches to Digital Research*. Emerald Group Publishing Limited 2014. 31–50.http://www.emeraldinsight.com/doi/abs/10.1108/S1042-319220140000013003 (accessed 22 Oct2015).

3 Verhulst S, Noveck B, Caplan R, *et al.* The Open Data Era in Health and Social Care. New York: : The GOVLAB (NYU) 2014. http://images.thegovlab.org/wordpress/wp-content/uploads/2014/10/nhs-full-report-21.pdf (accessed 10 May2015).

4 Willetts D. *Eight great technologies.* London, UK: : Policy Exchange 2013. http://www.policyexchange.org.uk/images/publications/eight%20great%20technologies.pdf

5 Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;**44**:827–36. doi:10.1093/ije/dyv098

6 Crossfield SSR, Clamp SE. Centralised Electronic Health Records Research Across Health Organisation Types. In: Fernández-Chimeno M, Fernandes PL, Alvarez S, *et al.*, eds. *Biomedical Engineering Systems and Technologies*. Berlin, Heidelberg: : Springer Berlin Heidelberg 2014. 394–406.http://link.springer.com/10.1007/978-3-662-44485-6_27 (accessed 10 Apr2016).

7 Denaxas SC, George J, Herrett E, *et al.* Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol* 2012;**41**:1625–38. doi:10.1093/ije/dys188

8 HSCIC. General Practice Trends in the UK to 2014. 2014.http://www.hscic.gov.uk/media/18173/General-Practice-Trends-in-the-UK-to-2014/pdf/gen-prac-trends-2014.pdf (accessed 14 Mar2016).

9 Walley T, Mantgani A. The UK General Practice Research Database. *Lancet Lond Engl* 1997;**350**:1097–9. doi:10.1016/S0140-6736(97)04248-7

10 Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. *Inform Prim Care* 2004;**12**:171–7.

11 Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research. *Inform Prim Care* 2004;**12**:49–50.

12 Williams T, van Staa T, Puri S, *et al.* Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther Adv Drug Saf* 2012;**3**:89–99. doi:10.1177/2042098611435911

13 Gnani S, Azeem M. A user's guide to data collected in primary care in England. Imperial College London: : Eastern Region Public Health Observatory (erpho) on behalf of the Association of Public Health Observatories 2006. https://www1.imperial.ac.uk/resources/579D8B09-C1C1-4026-A7BE-C3E936EE9567/ (accessed 15 Oct2015).

14 van Staa T-P, Dyson L, McCann G, *et al.* The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health Technol Assess* 2014;**18**. doi:10.3310/hta18430

15 Hall GC, Sauer B, Bourke A, *et al.* Guidelines for good database selection and use in pharmacoepidemiology research: GOOD DATABASE CONDUCT IN PHARMACOEPIDEMIOLOGY. *Pharmacoepidemiol Drug Saf* 2012;**21**:1–10. doi:10.1002/pds.2229

16 Smeeth L, Cook C, Fombonne E, *et al.* MMR vaccination and pervasive developmental disorders: a case-control study. *Lancet Lond Engl* 2004;**364**:963–9. doi:10.1016/S0140-6736(04)17020-7

17 Freemantle N, Marston L, Walters K, *et al.* Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013;**347**:f6409–f6409. doi:10.1136/bmj.f6409

18 van Staa T-P, Goldacre B, Gulliford M, *et al.* Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ* 2012;**344**:e55–e55. doi:10.1136/bmj.e55

19 CPRD. Research Papers. https://www.cprd.com/Bibliography/Researchpapers.asp

20 QResearch. QResearch Research articles. http://www.qresearch.org/SitePages/publications.aspx

21 IMS Health. Bibliography. http://www.epic-uk.org/bibliography/bibliography_01.shtml

22 Jacomy M. Table2Net MediaLab Tool. Sci.-Po Medialab. : http://tools. medialab.sciences-po.fr/table2net/

23 Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. 2009.https://gephi.github.io

24 Jacomy M, Venturini T, Heymann S, *et al.* ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* 2014;**9**:e98679. doi:10.1371/journal.pone.0098679

25 van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 2010;**84**:523–38. doi:10.1007/s11192-009-0146-3

26 van Eck NJ, Waltman L. Visualizing Bibliometric Networks. In: Ding Y, Rousseau R, Wolfram D, eds. *Measuring Scholarly Impact*. Cham: : Springer International Publishing 2014. 285–320.http://link.springer.com/10.1007/978-3-319-10377-8_13 (accessed 5 Jan2016).

27 Neimann AL, Shin DB, Wang X, *et al.* Prevalence of cardiovascular risk factors in patients with psoriasis. *J Am Acad Dermatol* 2006;**55**:829–35. doi:10.1016/j.jaad.2006.08.040

28 Smeeth L, Thomas SL, Hall AJ, *et al.* Risk of myocardial infarction and stroke after acute infection or vaccination. *N Engl J Med* 2004;**351**:2611–8. doi:10.1056/NEJMoa041747

29 Dial S, Delaney J a. C, Barkun AN, *et al.* Use of gastric acid-suppressive agents and the risk of community-acquired Clostridium difficile-associated disease. *JAMA* 2005;**294**:2989–95. doi:10.1001/jama.294.23.2989

30 Jick H, Jick SS, Gurewich V, *et al.* Risk of idiopathic cardiovascular death and nonfatal venous thromboembolism in women using oral contraceptives with differing progestagen components. *Lancet Lond Engl* 1995;**346**:1589–93.

31 Currie CJ, Poole CD, Gale E a. M. The influence of glucose-lowering therapies on cancer risk in type 2 diabetes. *Diabetologia* 2009;**52**:1766–77. doi:10.1007/s00125-009-1440-6

32 Yang Y-X, Lewis JD, Epstein S, *et al.* Long-term proton pump inhibitor therapy and risk of hip fracture. *JAMA* 2006;**296**:2947–53. doi:10.1001/jama.296.24.2947

33 Henry D, Lim LL, Garcia Rodriguez LA, *et al.* Variability in risk of gastrointestinal complications with individual non-steroidal anti-inflammatory drugs: results of a collaborative meta-analysis. *BMJ* 1996;**312**:1563–6.

34 Van Staa TP, Leufkens HGM, Abenhaim L, *et al.* Use of oral corticosteroids and risk of fractures. June, 2000. *J Bone Miner Res Off J Am Soc Bone Miner Res* 2005;**20**:1487–94; discussion 1486. doi:10.1359/jbmr.2005.20.8.1486

35 Gelfand JM, Neimann AL, Shin DB, *et al.* Risk of myocardial infarction in patients with psoriasis. *JAMA* 2006;**296**:1735–41. doi:10.1001/jama.296.14.1735

36 Jick H, Zornberg GL, Jick SS, *et al.* Statins and the risk of dementia. *Lancet Lond Engl* 2000;**356**:1627–31.

37 Dodd CN, Romio SA, Black S, *et al.* International collaboration to assess the risk of Guillain Barré Syndrome following Influenza A (H1N1) 2009 monovalent vaccines. *Vaccine* 2013;**31**:4448–58. doi:10.1016/j.vaccine.2013.06.032

38 CEIFE. Home. http://www.ceife.es/index.html# (accessed 15 Oct2015).

39 Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E* 2004;**69**. doi:10.1103/PhysRevE.69.026113

40 UCL. People. https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/staff

41 van Staa T-P. Professor Tjeerd Van Staa. http://www.population-health.manchester.ac.uk/staff/158950/

42 QResearch. QResearch Organisation. http://www.qresearch.org/SitePages/Organisation.aspx (accessed 15 Oct2015).

43 Lawson DH, Sherman V, Hollowell J. The General Practice Research Database. Scientific and Ethical Advisory Group. *QJM Mon J Assoc Physicians* 1998;**91**:445–52.

44 Boston University. The Clinical Practice Research Datalink. https://www.bu.edu/bcdsp/gprd/

45 Scopus. Scopus Content Coverage Guide. https://www.elsevier.com/__data/assets/pdf_file/0007/69451/sc_content-coverage-guide_july-2014.pdf

46 Chisholm J. The Read clinical classification. *BMJ* 1990;**300**:1092.

47 Azoulay P, Fons-Rosen C, Zivin JSG. Does Science Advance One Funeral at a Time? Cambridge, MA: : National Bureau of Economic Research 2015. http://www.nber.org/papers/w21788.pdf (accessed 5 Jan2016).

48 Margolis R, Derr L, Dunn M, *et al.* The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc* 2014;**21**:957–8. doi:10.1136/amiajnl-2014-002974

49 Krumholz HM. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. *Health Aff (Millwood)* 2014;**33**:1163–70. doi:10.1377/hlthaff.2014.0053

50 Ioannidis JPA, Greenland S, Hlatky MA, *et al.* Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet* 2014;**383**:166–75. doi:10.1016/S0140-6736(13)62227-8

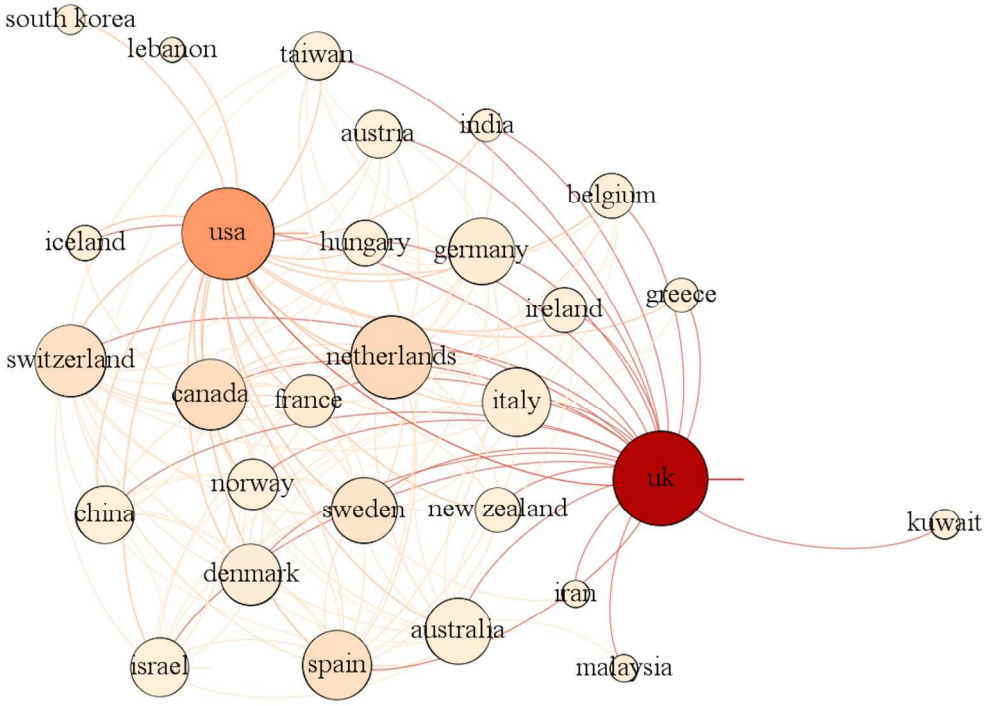

Fig 1. Network of contributing countries.
Fig 1. Network of contributing
190x137mm (300 x 300 DPI)

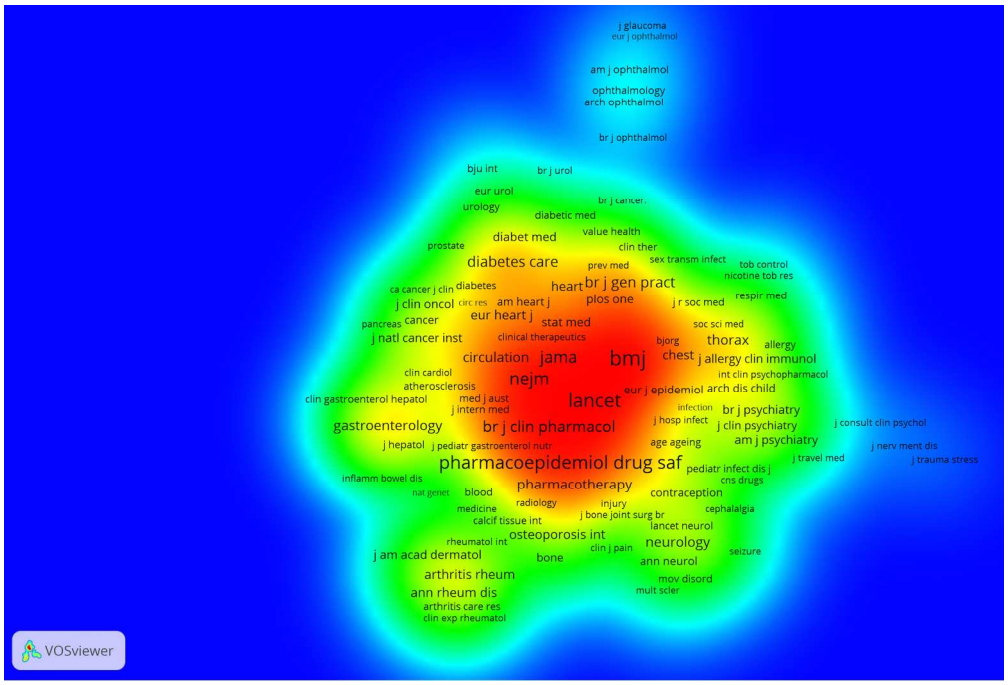

Fig 2. Journal co-citation analysis.
Fig 2. Journal co-citation ana
190x128mm (300 x 300 DPI)

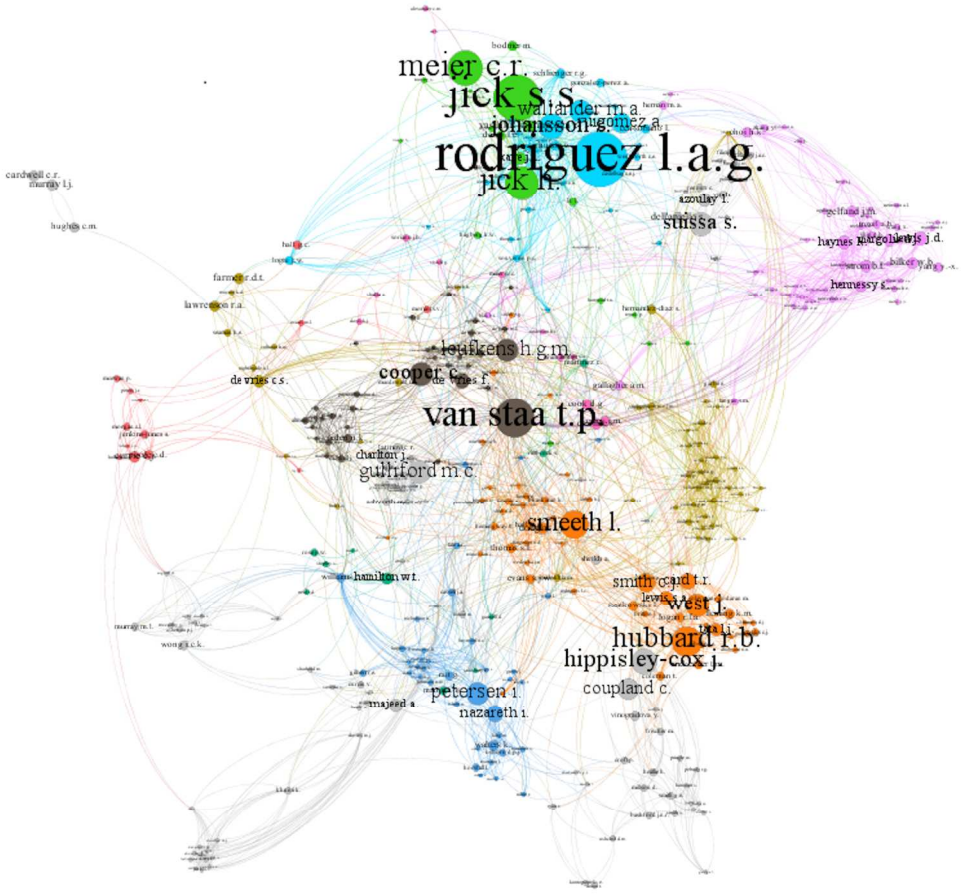

Fig 3. Clustered co-author network.
Fig 3. Clustered co-author net
190x176mm (300 x 300 DPI)

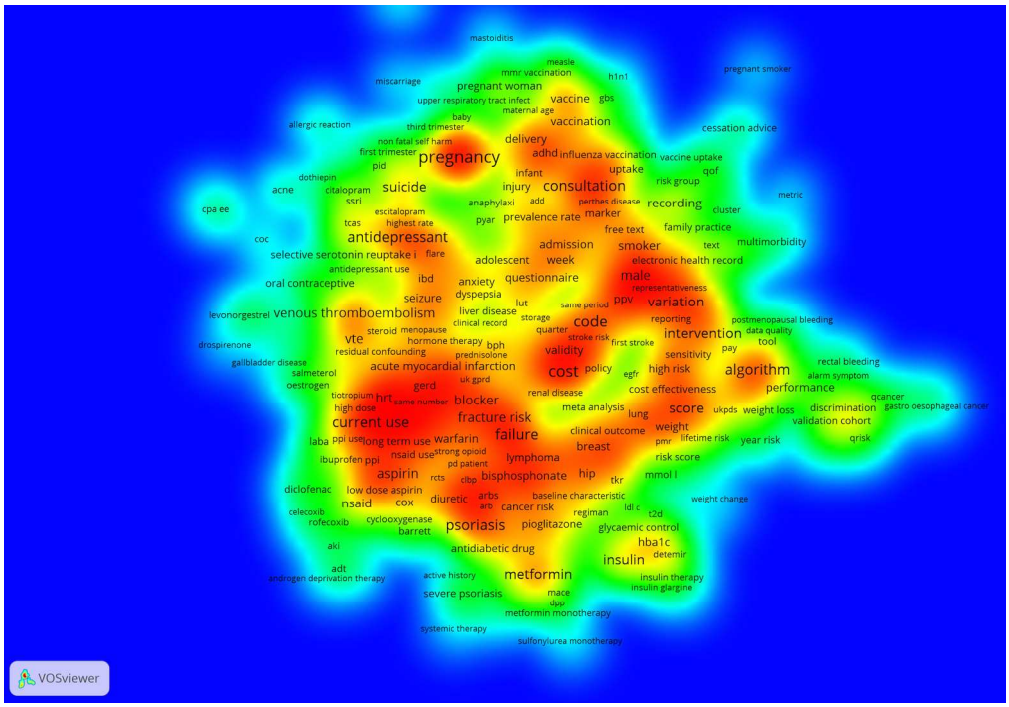

Fig 4. Term co-occurrence density map.
Fig 4. Term co-occurrence dens
190x132mm (300 x 300 DPI)

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	N/A
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	N/A
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	N/A
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	6

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	N/A
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	6
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	N/A
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	N/A
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	6-18
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/A
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	18
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	20
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	20
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	21

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

BMJ Open

The Evolution of Primary Care Databases in UK: A Scientometric Analysis of Research Output

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-012785.R1
Article Type:	Research
Date Submitted by the Author:	27-Jul-2016
Complete List of Authors:	Vezyridis, Paraskevas; University of Nottingham, Business School Timmons, Stephen; University of Nottingham, Business School
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Evidence based practice
Keywords:	scientometrics, PRIMARY CARE, electronic patient records, STATISTICS & RESEARCH METHODS

The Evolution of Primary Care Databases in UK: A Scientometric Analysis of Research Output

Paraskevas Vezyridis[1*], Stephen Timmons[1]

[1] Centre for Health Innovation, Leadership and Learning (CHILL), Nottingham University Business School, Jubilee Campus, Nottingham NG8 1BB, UK

* Corresponding author

E-mail: Paraskevas.Vezyridis@nottingham.ac.uk

Telephone: +44 115 846 6370

Fax: +44 115 846 6667

Abstract

Objective To identify publication and citation trends, most productive institutions and countries, top journals, most cited articles and authorship networks from articles that used and analysed data from primary care databases (CPRD, THIN, QResearch) of pseudonymised electronic health records in UK.

Methods Descriptive statistics and scientometric tools were used to analyse a SCOPUS dataset of 1891 articles. Open access software was used to extract networks from the dataset (Table2Net), visualise and analyse co-authorship networks of scholars and countries (Gephi) and, density maps (VOSviewer) of research topics co-occurrence and journal co-citation.

Results Research output increased overall at a yearly rate of 18.65%. While medicine is the main field of research, studies in more specialised areas include biochemistry and pharmacology. Researchers from UK, USA and Spanish institutions have published the most papers. Most of the journals that publish this type of research and most cited papers come from UK and USA. Authorship varied between 3-6 authors. Keyword analyses show that smoking, diabetes, cardiovascular diseases and mental illnesses, as well as medication that can treat such medical conditions, such as non-steroid anti-inflammatory agents, insulin and antidepressants constitute the main topics of research. Co-authorship network analyses show that lead scientists, directors or founders of these databases are, to various degrees, at the centre of clusters in this scientific community.

Conclusions There is a considerable increase of publications in primary care research from electronic health records. The UK has been well placed at the centre of an expanding global scientific community, facilitating international collaborations and bringing together international expertise in medicine, biochemical and pharmaceutical research.

Strengths and limitations of this study

- First study to perform a scientometric analysis of research output from primary care databases of electronic patient records.
- We analysed articles published from 1995 to 2015 in order to explore the historical breadth and growth of this type of research.
- The analysis is limited on articles and structured data retrieved from the Scopus database.
- Some latest articles and related citations might not have appeared in Scopus when the dataset was extracted.

Introduction

Big data (analytics) refers to the aggregation and interrogation of – high volume, high velocity, high variety – datasets so as to reveal new, non-obvious, information and patterns [1]. This field is advancing because of technological and scientific developments in information infrastructure and digitisation [2]. For governments, opening up the datasets states hold about their citizens is believed to have, through computational and algorithmic analyses, a disruptive and transformative effect on knowledge [3]. In the United Kingdom (UK), big (open) data has been at the forefront of research activity and policy-making. Termed as one of the *eight great technologies* [4], UK has embraced the big (open) data movement more than many other developed countries (e.g. US, Australia, France) [3]. One area of particular relevance to big data analytics is healthcare.

In UK, the National Health Service (NHS) is organised around primary care and, unless there is an accident or emergency, whenever citizens would like to use the NHS they have to go through their primary care physician, known in the UK as a General Practitioner (GP). From there, they can be referred to a specialist at a hospital if necessary. Secondary care clinicians can then feedback information to GPs. Since the vast majority of the population (98%) is registered with a general practice, GPs act not only as the main gatekeepers for the NHS but also as important custodians of a longitudinal electronic health record (EHR) [5]. There are now many ongoing primary care databases of anonymised patient records in UK that can be used for healthcare research. These population-based databases contain data originating from routine general practice. Some newly established databases and research platforms of linked EHRs include ResearchOne [6] and CALIBER [7]. While there are more than 9,600 general practices in UK that could potentially contribute data to these databases [8], it is usually 6-10% of these practices that do so.

Such databases are usually used for cross sectional surveys, case control or cohort studies and for epidemiological, drug safety, clinical and healthcare utilisation research purposes. They rely heavily on individual general practices voluntarily contributing data via the propriety clinical systems they use to maintain these patient records. The records are usually anonymised or pseudo-anonymised at source by allocating a unique number at each patient to allow for the updates of the records as well as for their linkage to other datasets, such as national mortality, national cancer registration and hospital records as well as with socioeconomic, ethnicity and environmental datasets. Access to these datasets is usually granted after scientific and ethics review and can be tailored to customer requirements. In this study, we examined the research output of three such databases that are well established in the research community and have contributed to a substantial number of

scientific studies and publications. These are the Clinical Practice Research Datalink (CPRD) [9], The Health Improvement Network (THIN) [10] and QResearch [11].

The CPRD (formerly known as the General Practice Research Database) is a not-for-profit research service funded by the NHS National Institute for Health Research (NIHR) and the Medicines and Healthcare products Regulatory Agency (MHRA). It is owned by the UK Department of Health and contains the records of 11 million patients (4.4 million active) from 674 general practices [5]. There is a service cost associated with the preparation of the requested data. Unlike the other services described below, CPRD does not extract data from a particular propriety clinical system. Any general practice can contribute data after a data sharing agreement with the software supplier. Importantly, it is the only database accessible online [12]. The THIN database contains the health records of 12 million patients from around 600 general practices that use the Vision clinical system by INPS. IMS Health can provide access to data, for example, via a yearly sub-license to an academic institution. THIN is the only database that can provide access to data for for-profit companies. QResearch is a research service located at the University of Nottingham. Its database contains the health records of 18 million patients from 1000 general practices that use the EMIS clinical system. Only academics employed by a UK university can have (in site) access only to a *sample* of the dataset (maximum 100,000 patients) that is sufficient to answer a specific research question or hypothesis. As this research service is not-for-profit and entirely self-funded there is usually a fee to be paid to cover the cost of the data extraction.

The strengths of these databases lie in their size, breadth, representativeness of the UK population, long-term follow-up and data quality [5]. They include good information on morbidity and lifestyle, prescribing, preventive care, current standards of care and inter-practice variation [13]. Since they are continually (and automatically) updated, they are ideal for researchers to discover and monitor healthcare trends as well as the effectiveness of new interventions and treatments, with minimum cost. They are increasingly linked to secondary care and mortality datasets. In contrast, their weaknesses include the fact that data are extracted from propriety clinical systems developed for patient management and not for healthcare research. There are issues of missing data (e.g. from healthy patients), variable definitions for diagnoses, and incomplete secondary care data (e.g. hospital admissions). Wider health data (e.g. treatment adherence, over the counter medication) and data about sub-populations (e.g. prisoners, homeless people, refugees, travellers) not captured adequately [5]. Information governance and informed consent procedures around the data sharing of EHRs for research are still considered complex [14]. These databases also require considerable clinical and scientific expertise, as well as technical capacity in data management to support research. When selecting a particular data resource for an observational

study, researchers have to consider several other factors, such as the population covered and its geographical distribution, data capture and latency, linkage with other resources, privacy and security, quality and validation [15].

Nonetheless, these databases are highly regarded within the research community since they have proved their value in helping researchers reach definitive answers in various healthcare debates of considerable public interest, particularly where other types of research have produced contradictory evidence. For example, in 2004 researchers from UK and Canada proved beyond doubt that measles-mumps-rubella (MMR) vaccination is not associated with autism in children [16]. In contrast to expensive, time-consuming, and unrepresentative (of the population) traditional randomised trials [17], large-scale and randomised observational (comparative) evaluations of treatments and medications are minimally obtrusive for clinicians and patients and can support faster turnaround times for pragmatic evidence useful in clinical practice [18].

The aim of this study was to perform a scientometric analysis of articles, published from 1995 to 2015, which have used data from at least one of these primary care databases. This empirical, semi-automated, method of quantitatively analysing a large number of publications provides a reliable and objective examination of the current status and trends as well as the structure and dynamics of this scientific field [19,20]. In this way, policy makers, research funding bodies but most importantly new researchers entering this field can have a general overview of its knowledge base and an indication of what kind of network features, research activities and topics of interest are driving it [20,21]. To the best of our knowledge, this is the first study of a systematic mapping of primary care databases research output.

Methods

In this study, Elsevier's Scopus database (www.scopus.com) was selected as the source of structured data on articles. This database covers more scientific articles than other databases (e.g. Thomson Reuters Web of Science) and has the advantage of providing advance export functionality of structured data including full citation information, abstracts, keywords and references. On October 30, 2015, we searched, using the *document search* functionality, for all articles containing the terms "General Practice Research Database (GPRD)" OR "Clinical Practice Research Datalink (CPRD)" OR "The Health Improvement Network (THIN)" OR "QResearch" in Article Title, Abstract and Keywords.

The results were then limited to *articles*, *articles in press*, *conference papers*, *reviews*, *book chapters* and *short surveys*. *Notes*, *letters*, *editorials* and *errata* were excluded from the analysis. From there, we compared the

resulted records with the bibliographic lists maintained by these databases [22–24] so as to include articles that could not be retrieved using the above search queries. Data cleansing included the removal of duplicate records and records that were missing essential information for the analysis (e.g. article title, journal). The fields of *authors*, *year*, *source title*, *affiliations*, *author keywords* and *document type* were used for the analysis.

The final bibliography retrieved from Scopus was imported to Table 2 Net [25] to extract networks of authors and contributing countries. It was then imported to Gephi [26] where the ForceAtlas 2 algorithm [27] was used to visualise the structural proximities for the communities of authors and contributing countries. The VOSviewer (version 1.6.3) [28] software was used to visualise bibliometric networks and densities [29] of frequent terms and journals. All other statistical analyses were performed using Microsoft Excel. We used the Journal Citation Reports (JCR) Science Edition 2014 to extract impact factor values for the identified journal titles.

Results

A total of 1,891 papers from 1995 to 2015 were included in this bibliometric and scientometric analysis. The results are presented below.

Publication and citation trends

The literature related to the 3 primary care databases in England increased gradually from 7 papers in 1995 to 171 in 2015 (Table 1). We estimated their Compound Annual Growth Rate (CAGR), for the years 1995-2014, to be 18.65%. The vast majority of papers were published in English across 425 different sources (16.76% CAGR for 1995-2014). In total, these papers have already been cited 73,929 times. There is, however, a small percentage of 1.16% (n=163) papers that have not yet been cited yet. The average citation per year is approximately 3.52.

Year	No. of Papers	%	No. of Citations	No. of Different Sources
2015	171	9.04	255	107
2014	214	11.31	1188	114
2013	203	10.73	1934	123
2012	175	9.25	3050	99
2011	148	7.82	3900	100
2010	129	6.82	5232	81

2009	126	6.66	5341	79
2008	115	6.08	4757	73
2007	96	5.07	5232	65
2006	74	3.91	5943	56
2005	81	4.28	5839	56
2004	73	3.86	6411	49
2003	46	2.43	2998	37
2002	52	2.74	3832	37
2001	51	2.69	3770	38
2000	51	2.69	6334	35
1999	26	1.37	1594	21
1998	29	1.53	2582	20
1997	17	0.89	1560	13
1996	7	0.37	1137	7
1995	7	0.37	1040	6
Total	**1891**	**100**	**73929**	**425**

Table 1. Distribution of scientific literature by year

We explored the distribution of publications by document type. This is presented in Table 2 to identify the preferences of scholars using these databases in their research to share knowledge. The vast majority of scholars prefer to publish the findings of their research through journals, particularly as original articles (96.5%).

Type	No. of Papers	%
Article	1825	96.5
Conference Paper	18	0.95
Book Chapter	4	0.21
Review	41	2.16
Short Survey	3	0.15

Table 2. Distribution of scientific literature by Document Type

Next, we analysed the distribution of papers based on the academic discipline in which they have been categorised by Scopus (Table 3) and by which each paper may be attributed to more than one subject area [30]. Since we analysed bibliographic data based on published research using primary care databases, it comes as no surprise that the vast majority of papers are under the *Medicine* category. There is, however, a considerable number of papers (~25%) under the categories *Biochemistry, Genetics and Molecular Biology* and *Pharmacology, Toxicology and Pharmaceutics*, which indicates an emphasis on the use of these databases for the study of medications. It also indicates the potential interest in these databases from the pharmaceutical sector.

Subject	No. of Papers	%
Medicine	1838	97.2
Biochemistry, Genetics and Molecular Biology	266	14.1
Pharmacology, Toxicology and Pharmaceutics	197	10.4
Neuroscience	78	4.1
Immunology and Microbiology	70	3.7
Agricultural and Biological Sciences	53	2.8
Nursing	44	2.3
Psychology	21	1.1
Arts and Humanities	13	0.7
Environmental Science	10	0.5

Table 3. Distribution of scientific literature by Discipline

Most Productive Institutions and Countries

For a deeper insight into contribution patterns and scientific impact, we first identified the top 10 institutions (by number of papers) authors have used as affiliation and then we analysed citation patterns (Table 4). We also analysed authors' affiliations based on the country of their institution. For this, each publication was assigned to its authors' respective affiliated countries so as to identify the network of multinational collaborations. The distribution of the top 10 contributing countries is presented in Table 5. Finally, we visualised the network of contributing countries using Gephi. We ended up with a network of 29 nodes and 175 edges (Fig 1). Each node

represents a country, while its size denotes the country's degree and the colour the number of papers. The thickness of interconnected lines (edges) denotes the number of co-authored papers between the countries.

Rank	Institution	No. of Papers	%	Total Citations	Median (IQR)	Country
1	University of Nottingham	266	14.06	11540	18 (6-44.75)	UK
2	Boston University	228	12.05	12328	21.5 (6-57)	USA
3	Centro Espanol de Investigacion Farmacoepidemiologica (CEIFE)	163	8.62	8493	26 (7-59.5)	Spain
4	University College London	156	8.25	5226	14 (4.75-37)	UK
5	London School of Hygiene & Tropical Medicine	124	6.55	3696	14 (4-40.5)	UK
6	University of Utrecht	118	6.24	5067	17.5 (4-48.75)	Netherlands
7	University of Pennsylvania	110	5.81	7508	24.5 (10-64.25)	USA
8	Medicines and Health Care Products Regulatory Agency	94	4.97	2623	14 (4-33.25)	UK
9	King's College London	92	4.86	2221	13 (4-32.75)	UK
10	University of Oxford	86	4.54	1806	9.5 (3-23.25)	UK

Table 4. Most productive institutions

Rank	Country	No. of Papers	%
1	UK	1202	63.56
2	USA	563	29.77
3	Spain	192	10.15
4	Netherlands	164	8.67
5	Switzerland	115	6.08
6	Canada	112	5.92
7	Sweden	106	5.60
8	Germany	64	3.38

9	France	51	2.69
10	Italy	36	1.90

Table 5. Top contributing countries

[Fig 1. Network of contributing countries.]

The majority of the most productive institutions are universities. Top universities include the University of Nottingham, Boston University, University College London (UCL), the London School of Hygiene & Tropical Medicine and the University of Utrecht. Apart from these academic institutions, a research unit in Spain (CEIFE) and the Medicines and Health Care Products Regulatory Agency (MHRA) in the UK are involved in primary care databases-based research. In this table we also report the medians along with the interquartile ranges. From this it seems that scholars from CEIFE, University of Pennsylvania and Boston University are co-authors in publications that are highly cited compared to the other institutions in this list. Switzerland, Canada, Sweden, Germany, France and Italy had no institution among the top ten list, although they were ranked among the top 10 productive countries. Most papers are published by scholars from UK (63.56%), followed by the United States (USA) and Spain. With the exception of USA and Canada, most of the productive countries are in Europe. What is particularly interesting in these two tables is that scholars in institutions from the USA and Spain produce not only a great number of publications but also publications that are widely recognised by this scientific community in terms of citations.

Taking into account the measurements of *Weighted Degree*, *Clustering*, *Eigen Vector Centrality* and *Betweeness Centrality* (Table 6) we observe that, once again the UK, followed by USA, is placed at the centre of this scientific community. With the highest degrees of all measurements, institutions from this country are the most well-connected and authoritative ones, facilitating the linking between institutions in other countries.

Rank	Country	Occurrences	Weighted Degree	Page Rank	Eigen Centrality	Closness Centrality	Betweeness Centrality
1	UK	5770	27.0	0.062	1.0	0.933	0.250
2	USA	2506	26.0	0.060	0.98	0.903	0.229
3	Netherlands	666	22.0	0.047	0.95	0.8	0.049
4	Canada	567	18.0	0.039	0.83	0.717	0.024
5	Switzerland	409	18.0	0.039	0.85	0.717	0.015
6	Italy	85	17.0	0.037	0.78	0.7	0.024

7	Spain	481	17.0	0.037	0.82	0.7	0.011
8	Australia	37	16.0	0.036	0.76	0.682	0.024
9	Sweden	315	16.0	0.036	0.77	0.682	0.023
10	Germany	121	16.0	0.036	0.71	0.682	0.020

Table 6. Top countries by centrality

Top Journals

In Table 7, we identify the top 10 journals where most research is published. Six of these journals are published by a UK publisher and the rest are published in the USA. The journal *Pharmacoepidemiology and Drug Safety* features at the top of list, followed by the *British Journal of Clinical Pharmacology* and *Pharmacotherapy*, which signifies the focus of research, produced from these primary care databases, on the safe use of medication. This focus can also be seen in Table 3, where (apart from medicine) most papers are published in the fields of *Biochemistry, Genetics and Molecular Biology* and *Pharmacology, Toxicology and Pharmaceutics* (Scopus classification), but also in Table 5, where most of the top cited papers refer to potential risk for particular complications/conditions from the use of specific medication. More specialised journals, like *Diabetes Care* and *Annals of the Rheumatic Diseases*, are also featured in this list, indicating a particular focus of research activity in specific spectrums of diseases.

Rank	Journal Name	No. of Papers	%	Publisher	Impact Factor	Open Access	Country
1	Pharmacoepidemiology and Drug Safety	115	6.08	Wiley	2.939	Hybrid	UK
2	British Medical Journal	100	5.28	BMJ Publishing Group	17.445	Full	UK
3	British Journal of General Practice	67	3.54	Royal College of General Practitioners	2.294	Full	UK
4	British Journal of Clinical Pharmacology	57	3.01	Wiley	3.878	Hybrid	UK
5	Plos One	51	2.69	Public Library of Science	3.234	Full	USA
6	BMJ Open	41	2.16	BMJ Publishing Group	2.271	Full	UK
7	Pharmacotherapy	34	1.79	Wiley	2.662	Hybrid	USA

8	Diabetes Care	30	1.58	American Diabetes Association	8.420	Hybrid	USA
9	Epidemiology	24	1.26	Wolters Kluwer	6.196	Hybrid	USA
10	Annals of the Rheumatic Diseases	23	1.21	BMJ Publishing Group	10.377	Hybrid	UK

Table 7. Top journals of published research

Four journals in this list are open access (BMJ, British Journal of General Practice, Plos One, BMJ Open), which greatly facilitates the sharing of knowledge without limitations. The BMJ enjoys widespread recognition of the high quality of its published studies, as indicated by the high impact factor. The rest are behind a pay wall but offer authors an open access option to publish their research (hybrid access). An extra column with the Impact Factors of these top 10 journals from the 2014 JCR was also added in the table.

What is also of particular importance in terms of scientific impact is that the 10 most cited papers identified below (see Table 8) have not been published in journals in this list. However, by performing a (full counting) analysis of co-citation links in VOSviewer (Fig 2) for journals cited in the Scopus dataset (minimum number of citations=10) we see that most of this list is represented here (blue - lowest density, red - highest density).

[Fig 2. Journal co-citation analysis.]

Most Cited Papers

Next, we focused on the top 10 papers [31–40] and calculated the total count of citations for each paper (Table 8) for the period 1995-2015. Citations totalled 7,194 (1.02%) of all citations in this dataset. It seems that these studies in dementia, psoriasis, fractures, cardiovascular diseases and gastrointestinal complications in relation to certain medications have been of great interest in this scientific community. The majority of the top 10 most cited papers (60%) are open access at the publisher's website and can be freely read by anyone.

Rank	Authors / Title	Year	Country	Journal	Impact Factor	Citations	Open Access
1	Jick H., Zornberg G.L., Jick S.S., Seshadri S., Drachman	2000	USA	*Lancet*	45.217	1322	No

	D.A. Statins and the risk of dementia						
2	Gelfand J.M., Neimann A.L., Shin D.B., Wang X., Margolis D.J., Troxel A.B. Risk of myocardial infarction in patients with psoriasis	2006	USA	*Journal of the American Medical Association*	35.289	854	Yes
3	Van Staa T.P., Leufkens H.G.M., Abenhaim L., Zhang B., Cooper C. Use of oral corticosteroids and risk of fractures	2000	UK	*Journal of Bone and Mineral Research*	6.832	796	Yes
4	Henry D., Lim L.L.-Y., Rodriguez L.A.G., Perez Gutthann S., Carson J.L., Griffin M., Savage R., Logan R., Moride Y., Hawkey C., Hill S., Fries J.T. Variability in risk of gastrointestinal complications with individual non-steroidal anti-inflammatory drugs: Results of a collaborative meta-analysis	1996	Australia, Spain, USA, New Zealand, UK	*British Medical Journal*	17.445	688	Yes
5	Yang Y.-X., Lewis J.D., Epstein S., Metz D.C. Long-term proton pump inhibitor therapy and risk of hip fracture	2006	USA	*Journal of the American Medical Association*	35.289	637	Yes
6	Currie C.J., Poole C.D., Gale E.A.M. The influence of glucose-lowering therapies on cancer risk in type 2 diabetes	2009	UK	*Diabetologia*	6.671	629	Yes
7	Jick H., Jick S.S., Gurewich V., Myers M.W., Vasilakis C. Risk of idiopathic cardiovascular death and nonfatal venous thromboembolism in women using oral contraceptives	1995	USA, UK	*Lancet*	45.217	612	No

	with differing progestagen components						
8	Dial S., Delaney J.A.C., Barkun A.N., Suissa S. Use of gastric acid-suppressive agents and the risk of community-acquired Clostridium difficile-associated disease	2005	Canada	*Journal of the American Medical Association*	35.289	582	Yes
9	Smeeth L., Thomas S.L., Hall A.J., Hubbard R., Farrington P., Vallance P. Risk of myocardial infarction and stroke after acute infection or vaccination	2004	UK	*New England Journal of Medicine*	55.873	546	Yes
10	Neimann A.L., Shin D.B., Wang X., Margolis D.J., Troxel A.B., Gelfand J.M. Prevalence of cardiovascular risk factors in patients with psoriasis	2006	USA	*Journal of the American Academy of Dermatology*	4.449	528	No

Table 8. Most cited papers

Of the 10 papers, 8 are single country papers, while none were singled authored. Again, the USA has a considerable presence in this list, producing papers that are highly cited. In addition, many of the highly productive authors identified below (Table 10) were also found in this list.

Authorship Patterns and Networks

Authorship distributions varied from single to a maximum of 155 authors – for a study about the feasibility of international collaboration to evaluate, based on a common protocol, the risk of Guillain-Barré syndrome following pH1N1 vaccination [41]. In total, there were 9385 authors involved in the 1981 papers during 1995-2015. In Table 9, we can see that more than three-quarters of all papers were published by 3 or more authors. Only 1% of papers were written by a single author, while 5 papers did not have any authorship details. Almost a quarter of all papers was published by 4 authors, which have been widely cited across this scientific community. This indicates the high degree of expert collaboration in this field, that is necessary in analysing millions of primary care records.

Rank	No. of Authors	No. of Papers	%	Citations	%

1	4	460	24.33	16,437	22.23
2	5	383	20.25	16,071	21.74
3	3	310	16.39	12,209	16.51
4	6	252	13.33	12,084	16.35
5	2	145	7.67	7,159	9.68
6	7	123	6.50	3,798	5.14
7	8	88	4.65	2,914	3.94
8	>11	48	2.54	1,792	2.42
9	9	28	1.48	617	0.83
10	10	25	1.32	412	0.56
11	1	19	1.00	436	0.59

Table 9. Co-authorship distribution

Table 10 provides the ranking of the top 10 scholars, first, in terms of research productivity based on the overall number of co-authored papers. While, generally, most scholars are from the UK, the Director of the Spanish Centre for Pharmacoepidemiologic Research (CEIFE) [42] is the scholar with the most published research from these primary care databases. Also, there are researchers in this field who do not necessarily come from the academic environment. The pharmaceutical sector is actively involved in knowledge production from electronic primary care records.

Rank	Author	Affiliation	Country	No. of Papers	Weighted Degree	Clustering	Eigen Centrality	Closness Centrality	Betweeness Centrality
1	Rodriguez, L.A.G.	Spanish Centre for Pharmacoepidemiologic Research (CEIFE)	Spain	166	82.0	0.082	0.346	0.337	0.073
2	Jick, S.S.	Boston University	USA	142	90.0	0.066	0.327	0.329	0.078
3	Van Staa, T.P.	University of Manchester	UK	115	144.0	0.063	1.0	0.403	0.217
4	Jick, H.	Boston University	USA	98	55.0	0.098	0.217	0.322	0.032
5	Meier, C.R.	University of Basel	Switzerland	94	54.0	0.116	0.232	0.294	0.014
6	Hubbard, R.	University of Nottingham	UK	86	82.0	0.092	0.441	0.359	0.067
7	Smeeth, L.	London School of Hygiene & Tropical Medicine	UK	79	89.0	0.101	0.578	0.366	0.067
8	Hippisley-Cox, J.	University of Nottingham	UK	72	57.0	0.144	0.244	0.325	0.065
9	Johansson, S.	AstraZeneca	Sweden	70	46.0	0.218	0.267	0.310	0.012

10	Cooper, C.	University of Southampton	UK	65	67.0	0.139	0.414	0.342	0.035
	West, J.	University of Nottingham	UK	65	42.0	0.213	0.230	0.310	0.012

Table 10. Most productive authors

Considering only those scholars who have co-authored at least 2 papers in this dataset, the analysis suggested a network (Fig 3) with 1261 nodes and 6186 edges. Here, each node represents an author, while its size denotes the number of author's papers. The interconnected lines (edges) denote the co-authored papers between those authors. For better visualisation, we limited the number of minimum degrees to 5 (maximum degrees=145). After a modularity measurement, to identify community structure [43], we observe some established collaborative teams (clusters with different colours) around specific and highly productive scholars in the analysis of data from primary care databases also found in Table 10. We also observe a new (blue) cluster around the lead statistician for THIN [44] – one of the 3 primary care databases studied.

[Fig 3. Clustered co-author network.]

Taking into account the measurements of *Weighted Degree*, *Clustering*, *Eigen Vector Centrality* and *Betweeness Centrality* (Table 10), results indicate a cluster placed at the centre of this scientific community. With the lowest degree of clustering and the highest degrees of all the other measurements, its prominent scholar is the most well-connected, facilitating, more than any other scholar, linking between other scientific clusters and scholars.

What is also particularly interesting is the fact that that some of the scholars (and their institutions) in this list are affiliated, to a certain extent, to these databases, having served or currently acting as their founders, directors, lead scientists or members of their scientific committee [45,46]. For example, lead scientists from the Boston Collaborative Drug Surveillance Program in Boston University were among the first who developed the technical and scientific capacity of these databases in pharmacoepidemiological research [47,48].

Research Topics

We conducted a keyword analysis to identify important topics of published research. For this, we first extracted from the bibliographic dataset 5,813 unique keywords as indexed by Scopus [30] to base our analysis on more

complete indexing information compared to authors' keywords. We retrieved the top 30 keywords for two specific categories: *medical conditions* and *medications/substances* (Tables 11 and 12). Next, we created a (full counting of) term co-occurrence density map in VOSviewer (Fig 4) by building a text corpus out of the title and abstract fields in the bibliographic dataset (minimum number of a term occurrence=10). In this way, we were able to identify topics that not only appear more frequently in the literature, but that were also strongly related to each other, forming clusters of topics. Blue colour indicates a low density of terms and red colour the highest density of terms. In many cases, the density map represents the frequency of indexed keywords in Tables 11 and 12. Clearly, smoking, diabetes, cardiovascular diseases, mental illnesses, psoriasis, obesity, pregnancy and cancer as well as medication and substances that can treat these medical conditions, such as aspirin, insulin, antidepressants and non-steroid anti-inflammatory agents (NSAID), have been of great interest for scholars using EHRs in primary care.

Rank	Keyword	Occurrences	Rank	Keyword	Occurrences
1	smoking	328	16	cardiovascular diseases	96
2	diabetes mellitus	223	17	myocardial infarction	94
3	hypertension	223	18	chronic obstructive lung disease	90
4	non insulin dependent diabetes mellitus	179	19	heart failure	86
5	depression	167	20	cerebrovascular accident	85
6	stroke	165	21	rheumatoid arthritis	83
7	asthma	158	22	epilepsy	81
8	diabetes mellitus, type 2	155	23	breast cancer	78
9	cancer risk	150	24	fracture	75
10	cardiovascular risk	147	25	psoriasis	75
11	cardiovascular disease	133	26	gastrointestinal hemorrhage	69
12	obesity	129	27	hip fracture	68
13	heart infarction	126	28	osteoporosis	68
14	pregnancy	125	29	colorectal cancer	65

15	ischemic heart disease	104	30	fractures, bone	65

Table 11. Top keywords: medical conditions

Rank	Keyword	Occurrences	Rank	Keyword	Occurrences
1	nonsteroid antiinflammatory agent	182	16	proton pump inhibitor	75
2	acetylsalicylic acid	154	17	warfarin	71
3	metformin	150	18	antidiabetic agent	69
4	corticosteroid	143	19	anticonvulsive agent	68
5	hydroxymethylglutaryl coenzyme a reductase inhibitor	138	20	serotonin uptake inhibitor	65
6	insulin	133	21	calcium channel blocking agent	64
7	antidepressant agent	124	22	anti-bacterial agents	62
8	beta adrenergic receptor blocking agent	108	23	hydroxymethylglutaryl-coa reductase inhibitors	62
9	hypoglycemic agents	91	24	oral antidiabetic agent	61
10	anti-inflammatory agents, non-steroidal	90	25	paracetamol	56
11	dipeptidyl carboxypeptidase inhibitor	88	26	diuretic agent	53
12	antihypertensive agent	82	27	ibuprofen	52
13	neuroleptic agent	80	28	simvastatin	52
14	antibiotic agent	77	29	tricyclic antidepressant agent	52
15	hemoglobin a1c	75	30	diclofenac	50

Table 12. Top keywords: medications/substances

[Fig 4. Term co-occurrence density map.]

Discussion

This study identified the leading institutions, countries, authors, journals and topics as well as their networks of published research that have used primary care databases in the UK to extract and analyse data from EHRs. There is a growing production of such papers which indicates the interest of a global and highly collaborative scientific community in this field and also the knowledge and insights that can be gained for healthcare improvement. Publication output increased from 7 papers in 1995 to 171 by October 2015 (18.65% CAGR for 1995-2014). It may be worth noting that by performing a similar, limited to the UK, search in Scopus for the same period and with the keyword "primary care" we found a 10.83% CAGR, which shows the increase in research conducted from these databases outstrips the field more generally. The vast majority of publications (96.5%) were journal articles. While this research field can be located, generally, in medicine, biochemical and pharmaceutical developments seem to be equally important aimed at addressing widespread medical conditions, such as diabetes, cardiovascular diseases, mental illnesses, smoking, obesity and cancer.

The UK has been well placed in this scientific field. This is partly due to the fact that there is now more than 30 years of data available in GP information systems [49]. The investment in developing primary care databases from EHRs for research purposes has placed the country at the centre of an network of collaborations across the globe, bringing together international expertise for the analysis of ever-expanding and increasingly interlinked clinical datasets from primary, secondary and tertiary care. Access to these datasets has also allowed researchers and institutions from other countries to develop their own programmes of research to answer important clinical questions. Six of the most productive institutions are located in the UK and 63.56% of publications were authored by scholars affiliated with this country, followed by the USA. Interestingly, the top institutions were not exclusively universities. Among them we can find a research unit in Spain (CEIFE) and the executive agency in UK that funds and runs the CPRD database (MHRA), while one of the most productive scholars is affiliated with a pharmaceutical company. This signifies the great interest of various actors, from academic, governmental and private sectors, in research with primary care databases.

The geographical trend can also be observed from the location of journals with the most published papers. Six of the top 10 journals are published in the UK, followed by the USA. The journals with the most papers published included *Pharmacoepidemiology and Drug Safety*, *BMJ*, *British Journal of General Practice* and *British Journal of Clinical Pharmacology*, which signifies the great interest of scholars in using data from EHRs for pharmaceutical research. This is partially because one of the oldest sets of routine information collected by GP practices in the UK and made available by these databases is drug histories [47]. Regarding restrictions on access

to research outputs, only 4 journals in this list are fully open access. This may limit access to knowledge to researchers and members of the public that cannot afford subscription costs. Interestingly, it is the more established journals in medicine, like *JAMA*, *Lancet* and *NEJM*, that have published some of the most cited papers in this bibliometric dataset and enjoy a high level of co-citation activity.

Keyword analyses show that smoking, diabetes, cardiovascular diseases, mental illnesses, psoriasis, obesity, pregnancy and cancer constitute the main topics of research activity using EHRs in primary care. Often, this research concentrates on developing algorithms to identify risk of occurrence of a particular disease. Researchers are also interested in investigating medications that can treat these medical conditions, such as aspirin, other non-steroid anti-inflammatory agents (NSAIDs), insulin and antidepressants.

For the vast majority of publications, authorship varied between 3-6 authors, indicating widely collaborative, international, efforts to promote research in this field. Co-authorship network analyses showed that the lead scientists, directors and founders of these databases were found, to various degrees, at the centre of clusters in this scientific community, highlighting their invaluable contribution to knowledge production. As Azoulay et al. [50] have demonstrated in their study about eminent researchers and the vitality of a field, the development of co-authorship networks and clusters of collaborators in newly established scientific domains might be useful to boost research productivity. Based on each database's data access requirements, their established researchers appear to have a fundamental role in facilitating and promoting international collaborations for more researchers, institutions and countries. Importantly, they have a clear and in-depth understanding of the kind of research activities these databases can support in terms of data quality, structure and EHR coding practices. As these databases are expected to open up in the future to more stakeholders from various disciplines around health and as universities prepare to incorporate training in data science skills (e.g. statistics, biomedical informatics, biology and medicine) [51] into their clinical curricula, so as to nurture the next generation of clinical investigators [52], these established researchers could promote quality, reliable and ethically appropriate scientific research [53] from complicated and highly contextual datasets.

Our study has the typical limitations of a scientometric study. We analysed articles published in a period of 20 years in order to explore the historical breadth and growth of research from electronic primary care records. However, this analysis is limited on structured data retrieved from one bibliometric database of peer-reviewed literature. Therefore, only articles published in journals in its index were analysed. Also, some of the latest articles and related citations might not have been retrieved at the time of the search, which might explain the decrease in the number of publications and citations particularly from 2010 onwards. It was beyond the scope of

this quantitative study to assess the scientific quality and the socio-economic impact of the large number of publications analysed here. These studies have deployed a range of study designs across many sub-fields of primary care research and with various research findings. Our main objective was restricted to assessing one aspect of academic impact and research quality, i.e. patterns and trends in research outputs [54]. Future research could focus on the wider academic and socio-economic impact of these studies by examining the relationship between publications, citation patterns and collaborations with the development of new scientific methods in the field or of new medical products and healthcare services.

In conclusion, output of primary care research from electronic health records has consistently increased since their development. The development of these databases in the UK has placed the country and affiliated academic institutions at the centre of an expanding global scientific community, facilitating international collaborations and bringing together international expertise in medicine, biochemical and pharmaceutical research.

Contribution Statement

[Author] and [Author] conceived and designed the study. [Author] collected and analysed the data. [Author] and [Author] interpreted the findings, drafted the manuscript and approved the final manuscript for submission.

Competing Interests

The authors declare no competing interests.

Funding

[Author] is supported by a Marie Skłodowska-Curie Individual Fellowship from European Commission (xxxx-IF-xxxxxx). [Author] is partially supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care [East Midlands]. The views expressed are those of the authors and not necessarily those of the EC, NHS, the NIHR or the Department of Health.

Data Sharing

No additional data available.

References

1 Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: ethical issues. London, UK: 2015.

2 Goffey A, Pettinger L, Speed E. Politics, Policy and Privatisation in the Everyday Experience of Big Data in the NHS. In: Hand M, Hillyard S, eds. *Big Data? Qualitative Approaches to Digital Research*. Emerald Group Publishing Limited 2014. 31–50.http://www.emeraldinsight.com/doi/abs/10.1108/S1042-319220140000013003 (accessed 22 Oct2015).

3 Verhulst S, Noveck B, Caplan R, *et al.* The Open Data Era in Health and Social Care. New York: : The GOVLAB (NYU) 2014. http://images.thegovlab.org/wordpress/wp-content/uploads/2014/10/nhs-full-report-21.pdf (accessed 10 May2015).

4 Willetts D. *Eight great technologies.* London, UK: : Policy Exchange 2013. http://www.policyexchange.org.uk/images/publications/eight%20great%20technologies.pdf

5 Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;**44**:827–36. doi:10.1093/ije/dyv098

6 Crossfield SSR, Clamp SE. Centralised Electronic Health Records Research Across Health Organisation Types. In: Fernández-Chimeno M, Fernandes PL, Alvarez S, *et al.*, eds. *Biomedical Engineering Systems and Technologies*. Berlin, Heidelberg: : Springer Berlin Heidelberg 2014. 394–406.http://link.springer.com/10.1007/978-3-662-44485-6_27 (accessed 10 Apr2016).

7 Denaxas SC, George J, Herrett E, *et al.* Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol* 2012;**41**:1625–38. doi:10.1093/ije/dys188

8 HSCIC. General Practice Trends in the UK to 2014. 2014.http://www.hscic.gov.uk/media/18173/General-Practice-Trends-in-the-UK-to-2014/pdf/gen-prac-trends-2014.pdf (accessed 14 Mar2016).

9 Walley T, Mantgani A. The UK General Practice Research Database. *Lancet Lond Engl* 1997;**350**:1097–9. doi:10.1016/S0140-6736(97)04248-7

10 Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. *Inform Prim Care* 2004;**12**:171–7.

11 Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research. *Inform Prim Care* 2004;**12**:49–50.

12 Williams T, van Staa T, Puri S, *et al.* Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther Adv Drug Saf* 2012;**3**:89–99. doi:10.1177/2042098611435911

13 Gnani S, Azeem M. A user's guide to data collected in primary care in England. Imperial College London: : Eastern Region Public Health Observatory (erpho) on behalf of the Association of Public Health Observatories 2006. https://www1.imperial.ac.uk/resources/579D8B09-C1C1-4026-A7BE-C3E936EE9567/ (accessed 15 Oct2015).

14 van Staa T-P, Dyson L, McCann G, *et al.* The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health Technol Assess* 2014;**18**. doi:10.3310/hta18430

15 Hall GC, Sauer B, Bourke A, *et al.* Guidelines for good database selection and use in pharmacoepidemiology research: GOOD DATABASE CONDUCT IN PHARMACOEPIDEMIOLOGY. *Pharmacoepidemiol Drug Saf* 2012;**21**:1–10. doi:10.1002/pds.2229

16 Smeeth L, Cook C, Fombonne E, *et al.* MMR vaccination and pervasive developmental disorders: a case-control study. *Lancet Lond Engl* 2004;**364**:963–9. doi:10.1016/S0140-6736(04)17020-7

17 Freemantle N, Marston L, Walters K, *et al.* Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013;**347**:f6409–f6409. doi:10.1136/bmj.f6409

18 van Staa T-P, Goldacre B, Gulliford M, *et al.* Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ* 2012;**344**:e55–e55. doi:10.1136/bmj.e55

19 Shahram F, Jamshidi A-R, Hirbod-Mobarakeh A, *et al.* Scientometric analysis and mapping of scientific articles on Behcet's disease. *Int J Rheum Dis* 2013;**16**:185–92. doi:10.1111/1756-185X.12087

20 Heilig L, Vob S. A Scientometric Analysis of Cloud Computing Literature. *IEEE Trans Cloud Comput* 2014;**2**:266–78. doi:10.1109/TCC.2014.2321168

21 Uuskula A, Toompere K, Laisaar KT, *et al.* HIV research productivity and structural factors associated with HIV research output in European Union countries: a bibliometric analysis. *BMJ Open* 2015;**5**:e006591–e006591. doi:10.1136/bmjopen-2014-006591

22 CPRD. Research Papers. https://www.cprd.com/Bibliography/Researchpapers.asp

23 QResearch. QResearch Research articles. http://www.qresearch.org/SitePages/publications.aspx

24 IMS Health. Bibliography. http://www.epic-uk.org/bibliography/bibliography_01.shtml

25 Jacomy M. Table2Net MediaLab Tool. Sci.-Po Medialab. : http://tools. medialab.sciences-po.fr/table2net/

26 Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. 2009.https://gephi.github.io

27 Jacomy M, Venturini T, Heymann S, *et al.* ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* 2014;**9**:e98679. doi:10.1371/journal.pone.0098679

28 van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 2010;**84**:523–38. doi:10.1007/s11192-009-0146-3

29 van Eck NJ, Waltman L. Visualizing Bibliometric Networks. In: Ding Y, Rousseau R, Wolfram D, eds. *Measuring Scholarly Impact*. Cham: : Springer International Publishing 2014. 285–320.http://link.springer.com/10.1007/978-3-319-10377-8_13 (accessed 5 Jan2016).

30 Scopus. Scopus Content Coverage Guide. Elsevier B.V. 2016. https://www.elsevier.com/__data/assets/pdf_file/0007/69451/scopus_content_coverage_guide.pdf (accessed 10 Mar2016).

31 Neimann AL, Shin DB, Wang X, *et al.* Prevalence of cardiovascular risk factors in patients with psoriasis. *J Am Acad Dermatol* 2006;**55**:829–35. doi:10.1016/j.jaad.2006.08.040

32 Smeeth L, Thomas SL, Hall AJ, *et al.* Risk of myocardial infarction and stroke after acute infection or vaccination. *N Engl J Med* 2004;**351**:2611–8. doi:10.1056/NEJMoa041747

33 Dial S, Delaney J a. C, Barkun AN, *et al.* Use of gastric acid-suppressive agents and the risk of community-acquired Clostridium difficile-associated disease. *JAMA* 2005;**294**:2989–95. doi:10.1001/jama.294.23.2989

34 Jick H, Jick SS, Gurewich V, *et al.* Risk of idiopathic cardiovascular death and nonfatal venous thromboembolism in women using oral contraceptives with differing progestagen components. *Lancet Lond Engl* 1995;**346**:1589–93.

35 Currie CJ, Poole CD, Gale E a. M. The influence of glucose-lowering therapies on cancer risk in type 2 diabetes. *Diabetologia* 2009;**52**:1766–77. doi:10.1007/s00125-009-1440-6

36 Yang Y-X, Lewis JD, Epstein S, *et al.* Long-term proton pump inhibitor therapy and risk of hip fracture. *JAMA* 2006;**296**:2947–53. doi:10.1001/jama.296.24.2947

37 Henry D, Lim LL, Garcia Rodriguez LA, *et al.* Variability in risk of gastrointestinal complications with individual non-steroidal anti-inflammatory drugs: results of a collaborative meta-analysis. *BMJ* 1996;**312**:1563–6.

38 Van Staa TP, Leufkens HGM, Abenhaim L, *et al.* Use of oral corticosteroids and risk of fractures. June, 2000. *J Bone Miner Res Off J Am Soc Bone Miner Res* 2005;**20**:1487–1494; discussion 1486. doi:10.1359/jbmr.2005.20.8.1486

39 Gelfand JM, Neimann AL, Shin DB, *et al.* Risk of myocardial infarction in patients with psoriasis. *JAMA* 2006;**296**:1735–41. doi:10.1001/jama.296.14.1735

40 Jick H, Zornberg GL, Jick SS, *et al.* Statins and the risk of dementia. *Lancet Lond Engl* 2000;**356**:1627–31.

41 Dodd CN, Romio SA, Black S, *et al.* International collaboration to assess the risk of Guillain Barré Syndrome following Influenza A (H1N1) 2009 monovalent vaccines. *Vaccine* 2013;**31**:4448–58. doi:10.1016/j.vaccine.2013.06.032

42 CEIFE. Home. http://www.ceife.es/index.html# (accessed 15 Oct2015).

43 Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E* 2004;**69**. doi:10.1103/PhysRevE.69.026113

44 UCL. People. https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/staff

45 van Staa T-P. Professor Tjeerd Van Staa. http://www.population-health.manchester.ac.uk/staff/158950/

46 QResearch. QResearch Organisation. http://www.qresearch.org/SitePages/Organisation.aspx (accessed 15 Oct2015).

47 Lawson DH, Sherman V, Hollowell J. The General Practice Research Database. Scientific and Ethical Advisory Group. *QJM Mon J Assoc Physicians* 1998;**91**:445–52.

48 Boston University. The Clinical Practice Research Datalink. https://www.bu.edu/bcdsp/gprd/

49 Chisholm J. The Read clinical classification. *BMJ* 1990;**300**:1092.

50 Azoulay P, Fons-Rosen C, Zivin JSG. Does Science Advance One Funeral at a Time? Cambridge, MA: : National Bureau of Economic Research 2015. http://www.nber.org/papers/w21788.pdf (accessed 5 Jan2016).

51 Margolis R, Derr L, Dunn M, *et al.* The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc* 2014;**21**:957–8. doi:10.1136/amiajnl-2014-002974

52 Krumholz HM. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. *Health Aff (Millwood)* 2014;**33**:1163–70. doi:10.1377/hlthaff.2014.0053

53 Ioannidis JPA, Greenland S, Hlatky MA, *et al.* Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet* 2014;**383**:166–75. doi:10.1016/S0140-6736(13)62227-8

54 Penfield T, Baker MJ, Scoble R, *et al.* Assessment, evaluations, and definitions of research impact: A review. *Res Eval* 2014;**23**:21–32. doi:10.1093/reseval/rvt021

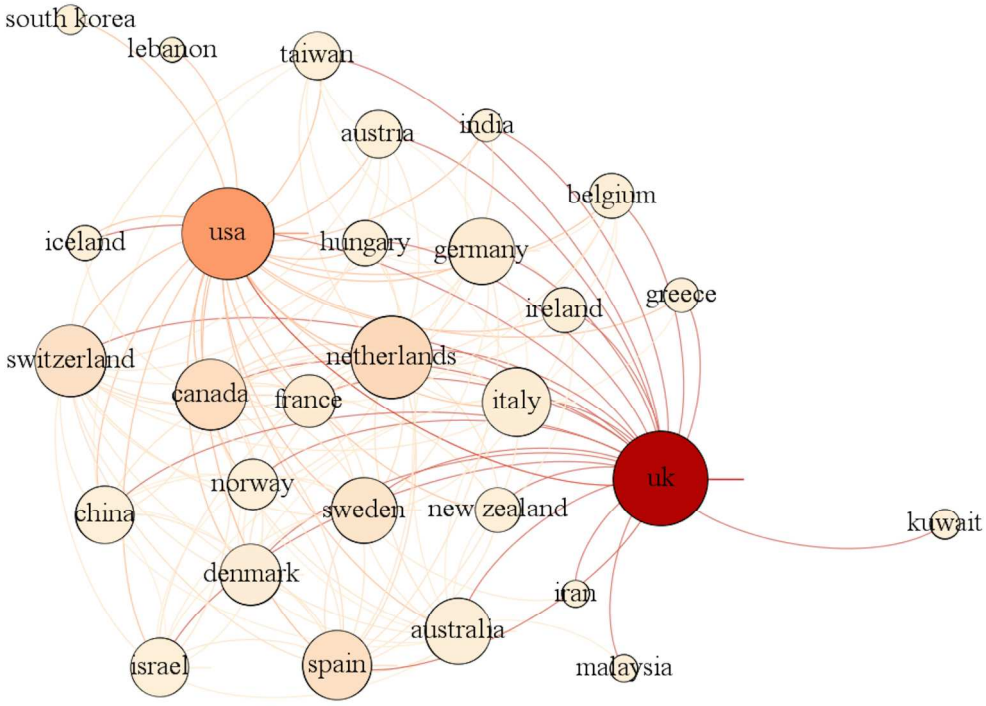

Fig 1. Network of contributing countries.
Fig 1. Network of contributing
190x137mm (300 x 300 DPI)

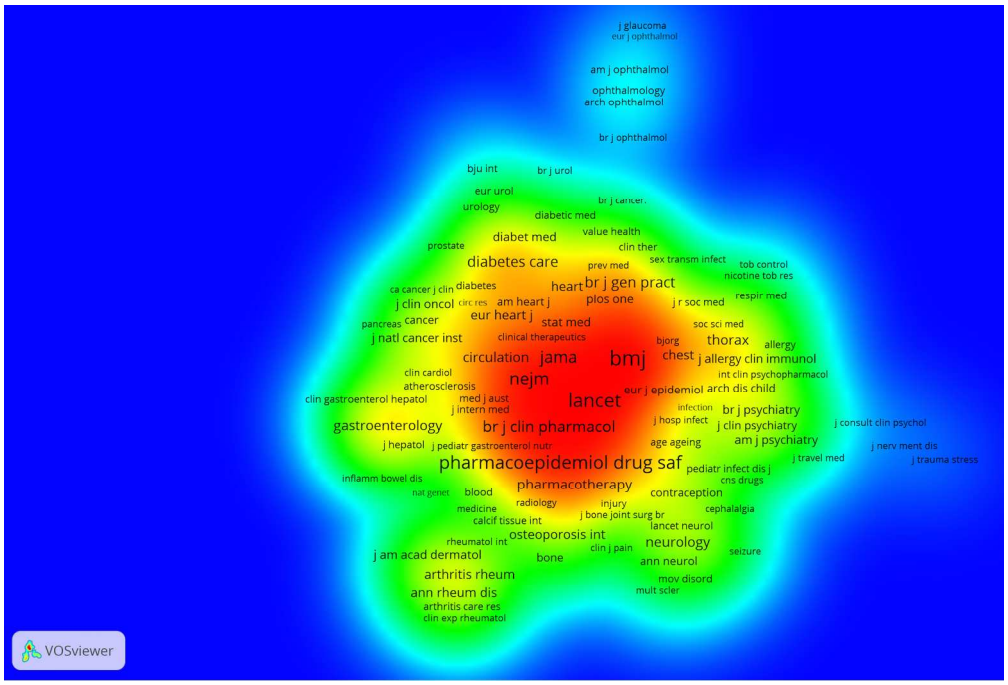

Fig 2. Journal co-citation analysis.
Fig 2. Journal co-citation ana
190x128mm (300 x 300 DPI)

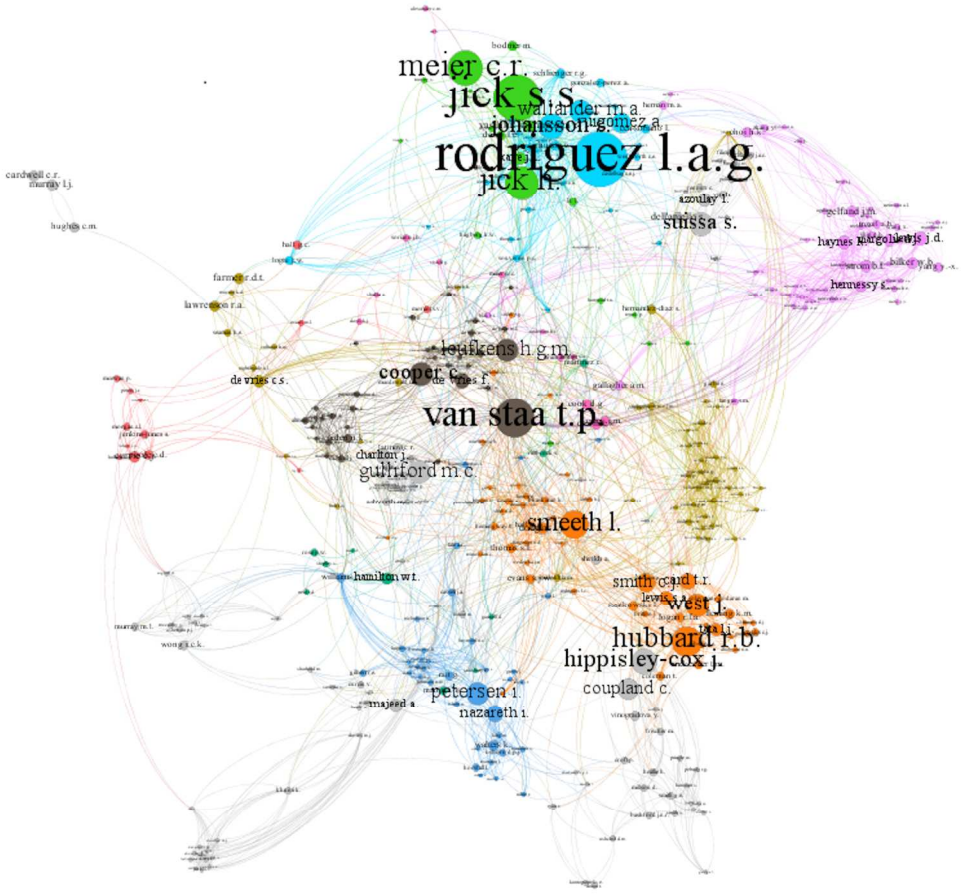

Fig 3. Clustered co-author network.
Fig 3. Clustered co-author net
190x176mm (300 x 300 DPI)

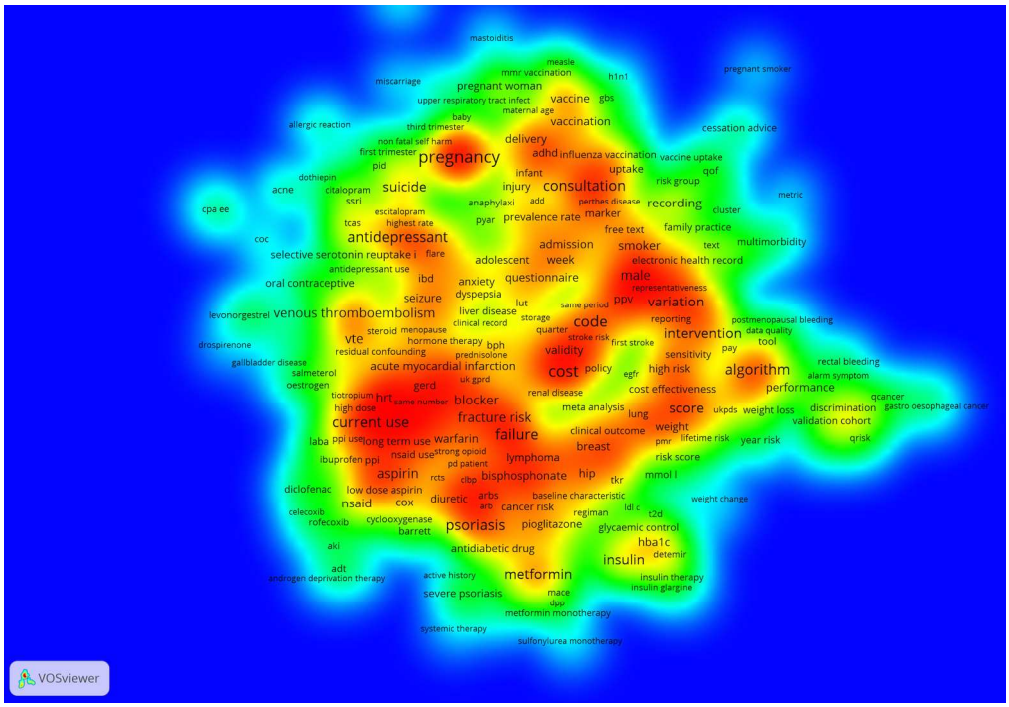

Fig 4. Term co-occurrence density map.
Fig 4. Term co-occurrence dens
190x132mm (300 x 300 DPI)

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	N/A
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	N/A
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	N/A
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	6

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	N/A
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	6
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	N/A
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	N/A
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	6-18
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/A
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	18
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	20
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	20
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	21

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

BMJ Open

Evolution of primary care databases in UK: a scientometric analysis of research output

Paraskevas Vezyridis and Stephen Timmons

BMJ Open 2016 6:
doi: 10.1136/bmjopen-2016-012785