The use of UK administrative data

An analysis of publications identified within the Dimensions system

September 2020

Economic and Social Research Council

The use of administrative data

Who uses the UK's administrative data, and what do they use it for? This is a simple question to ask, and an important one, but it is hard to answer. This analysis aims to shine some light on one aspect of that answer: the scientific publications that researchers produce as a result of their use of administrative data.

The aim of this analysis

This analysis is not an attempt to quantify the extent of use of administrative data, to assess the performance of the area, or somehow to rate the individuals who work with administrative data. It has the more limited aim of identifying and understanding usage of administrative data in primarily academic research and the community of researchers who use it. It does this through the lens of bibliometric information.

Bibliometric analysis of publications is a well-established, if problematic, way of understanding patterns of research and research impact.[1] UKRI is a signatory to DORA[2] and is committed to the responsible use of metrics and bibliometric data in general.

The data behind this analysis

In the absence of a comprehensive record of every publication produced which uses UK administrative data we have identified a list of key datasets[3] and searched the Dimensions system[4] for publications which reference them in their titles or abstracts. Publications which do not reference their use of a source of administrative data, which use a name variant not included in the search terms, or which reference their use of administrative data in another part of the publication will not have been included.

These are important limitations. The gaps in the data are significant enough that it cannot reasonably be described as a comprehensive analysis of the use of UK administrative data and we are not presenting it as such[5]. Instead it should be thought of as an illustrative and qualitative look at some of what is being done with some data of that kind. (It is worth highlighting that the use of data in ways which do not result in a research publication – for example in government planning – will be entirely absent, yet these uses may result in the greatest real-world impacts.)

The query identified articles indexed in Dimensions which were published in the three years 2017 to 2019 and which had at least one author with a UK affiliation. This is not long enough a period to produce sensible trends and so no reference is made to changes over time. After some data cleaning, just under 2,000 publications which indicated the use of administrative data were included.

Key findings

A large body of researchers have used administrative data in their work, and the resulting publications are well-received by other researchers

[1] See https://responsiblemetrics.org/the-metric-tide/ for a good summary of the issues.
[2] https://www.ukri.org/news/ukri-signs-san-francisco-declaration-of-research-assessment/
[3] These are listed in an annex.
[4] https://www.dimensions.ai/. UKRI has access to a private instance of Dimensions which contains more information than that used here. However, this analysis uses only publicly available data.
[5] The original underlying motivation for this analysis was to identify uses and users of administrative datasets that are available through ADR UK specifically, meaning that only those sets are used to create search terms. It is an analysis of the use of administrative data only in as far as the set of ADR UK datasets covers the area.

Around 5,300 unique authors contributed to the 2,000 publications.[6] 3,700 of these authors had at least one known UK affiliation and, not mutually exclusively, 1,300 had at least one known overseas affiliation.

These authors' publications tend to have a quite a high Field Citation Ratio (FCR; a measure of the relative extent of citation of that publication[7]) of around 3.5, meaning that each of them is cited roughly three and a half times as frequently as publications in the same area and of comparable age. This measurement is complicated and should be taken as indicative only, as the comparator group of publications is ill-defined.

Collaboration in research which uses administrative data is widespread, but rather focused.

In recent years, UK-affiliated authors of academic publications which make some use of administrative data have tended to favour collaboration with the USA and Australia. There are also notable, but much lower, levels of collaboration with co-authors in the Netherlands, Ireland and Sweden (Figure 1.)

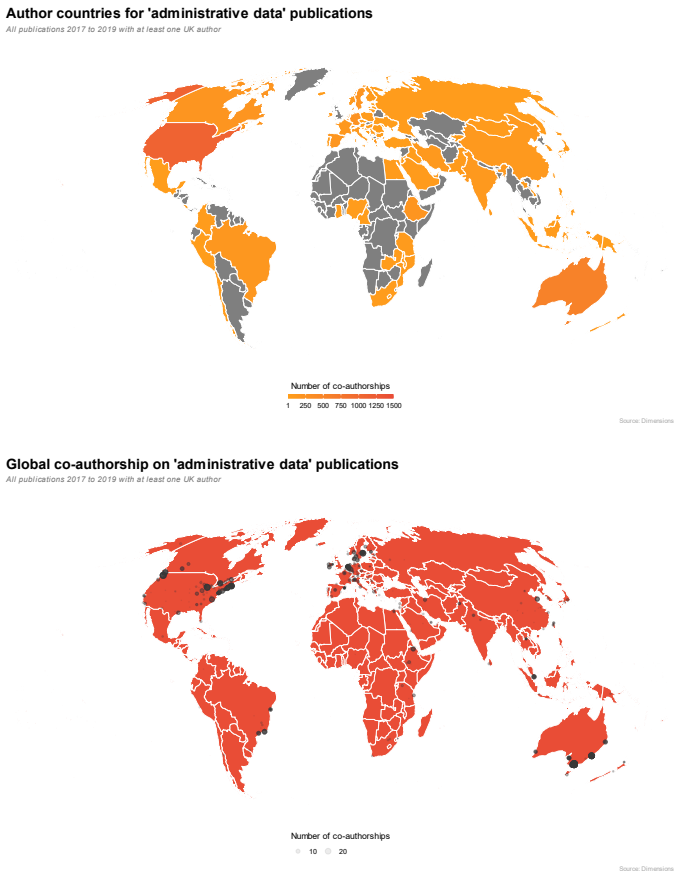

Figure 1: (top) national counts of total instances of co-authorship on 'administrative data' publications and (bottom) organisational count and location of those instances, for publications with at least one UK author affiliation in 2017 to 2019.

[6] The set of authors was small enough that it was possible to manually disambiguate all authors, meaning that this count is unlikely to be far from the true count, for this particular data set at least
[7] FCRs are not calculated for publications less than two years old, so any statement here that refers to them relates to publications from 2017 and 2018 only.

In each year, around 10% of publications in the data had at least one author from the USA, and about 5% had at least one author from Australia. While internationally-collaborative publications tended to have higher FCRs than did those with UK-only authors (median FCR 3.5 vs 3.3) the difference is not significant (p = .65)[8].

About three quarters of instances of authorship represent a UK-based affiliation and about 70% of publications have only UK authors. This suggests an internationalised flavour to research underpinned by the use of the UK's administrative data, but one which still reflects the source of the data.

The more authors that a publication has, the more opportunities there are for that publication to include at least one overseas author. And it is rare for a publication to have just one or two authors (Figure 2.)

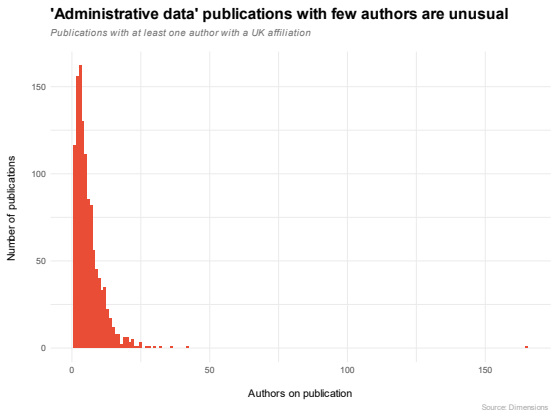

Figure 2: author counts on 'administrative data' publications 2017 to 2019.

The mean, median and modal average numbers of authors on these publications were 6.1, 5 and 3 respectively.

UK-based research activity using administrative data is widespread and a few large organisations are prominent in the field

Six UK universities – Edinburgh, Imperial, KCL, Oxford, UCL and Warwick – sit at the heart of a network of organisations hosting the authors of administrative data publications. About 20 other UK HEIs also feature prominently (Figure 3.)

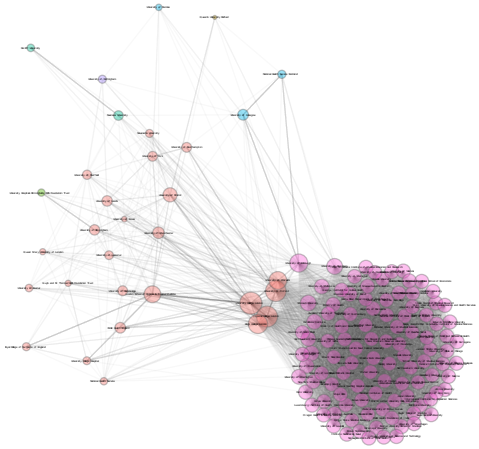

Figure 3: institutional author affiliation network for 'administrative data' publications 2017 to 2019. Organisations scaled by counts of authorships and coloured by algorithmically-assigned community membership.

A number of NHS trusts also host authors of publications which make use of administrative data, suggesting a strong healthcare relevance to much of the body of knowledge being created. Many overseas organisations, often appearing in the data at least as frequently as some UK collaborators, are also present in the network.

[8] Based on a permutation test with 10,000 replications. This p value means that were there no association between the international status of the authorship of a publication and its FCR, we would expect to see a difference in median FCR between the two groups at least as large as that actually observed about two thirds of the time. This is frequent enough that we can reasonably conclude that the result actually found has no meaningful implications.

The network of co-authorship created by publications which use the UK's administrative data is extensive and well-connected

As with the organisational network already shown, co-authorships can also be represented and interpreted as a network of individuals linked when they co-author a research publication.

Across this conceptual network, 60% of authorship instances are the only instance of that author in the data, and 80% of authors appear only once. Most individual authors publish relatively infrequently, but a few are much more prolific.

Slightly more than half of all the authors in the data (nearly 2,900 in all,) and 60% of all the publications, are linked in the same core group of authors and publications resulting from the use of UK administrative data[9].

This main component of the co-authorship network can be simplified to include only those authors whose presence in it most strongly connects all its members. A representation of this reduced network of highly-connective authors, in which individual authors are coloured on the basis of their membership of a community assigned algorithmically, is shown in Figure 4. The range of colours suggests a diverse set of focus areas across these leading authors.

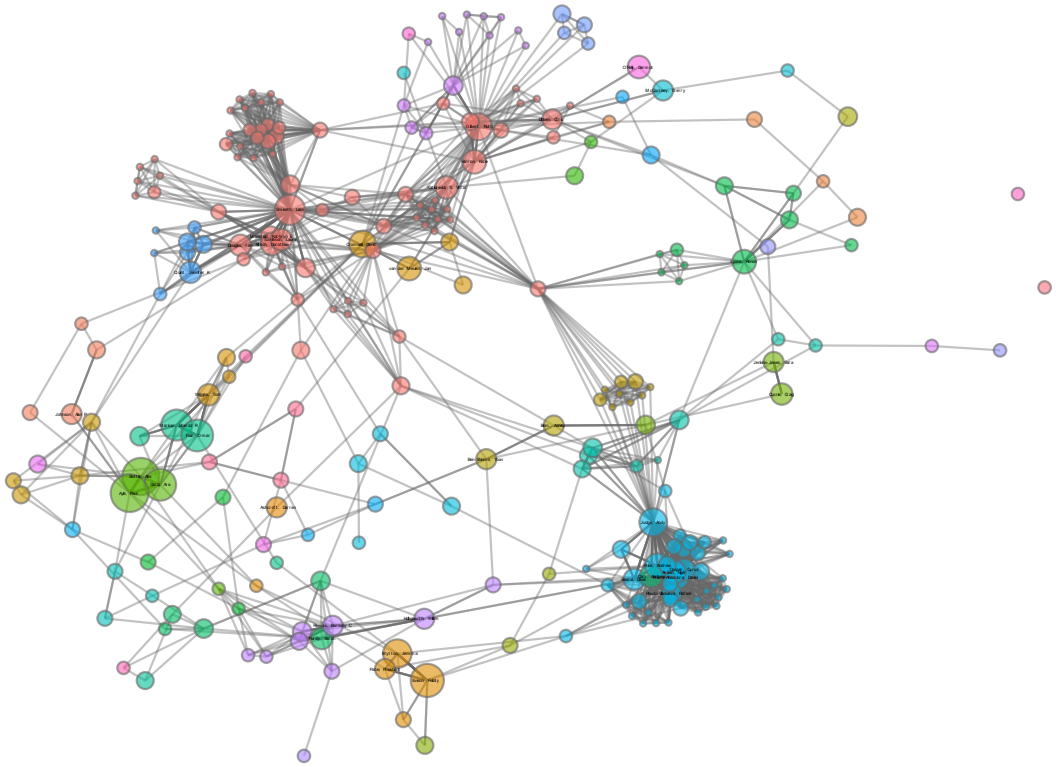

Figure 4: core network of authors on 'administrative data' publications with at least one UK-affiliated author in 2017 to 2019. Membership defined on the basis of the effect that deletion of each node has on the mean distance between nodes in the main graph component of the full network. Only the top 9% of authors by this measure are included. Nodes are coloured on the basis of communities identified with the walktrap algorithm, are labelled directly for authors with more than 5 authorship instances and are scaled to reflect that count.

[9] Based on similar analyses of Dimensions data from other subject areas it seems that this situation, in which most authors have a connection with most other authors, appears to be quite common.

The set of publications itself contains an extensive list of citations to previous work. Of the top 100 authors cited most frequently in publications in the dataset, about 50% are authors of publications in the dataset itself. Even allowing for an inevitable tendency to self-citation, the authors of the publications identified here appear to be contributing to the development of the field rather than being reliant on others' work.

About one publication in eight in the dataset acknowledges ESRC support[10]. Although the median FCR for publications known to have relied in some way on ESRC funding (2.9) is nearly twice that of publications not known to have received ESRC support (1.5) the difference is not significant (p = .67).

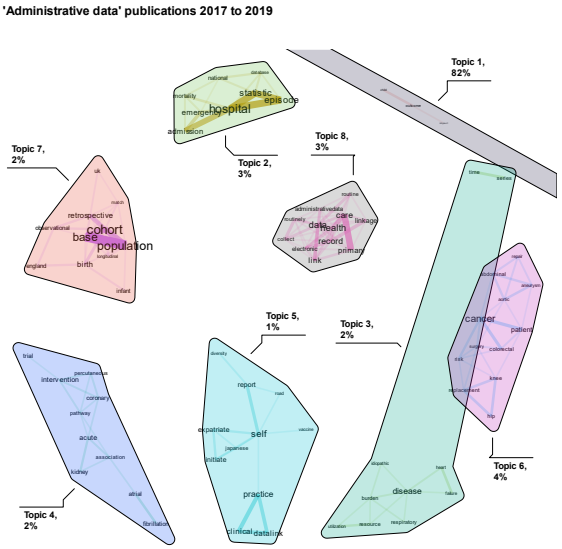

Figure 5:representation of a biterm topic model for titles of UK-affiliated 'administrative data' publications 2017 to 2019. Some key words in each topic are indicated directly.

Health research is a key focus for users of the UK's administrative data

A biterm topic model developed with these 2,000 publication titles again indicates that authors using UK administrative data often focus on issues relating to various aspects of individual and collective health[11]. Other topics, such as education, wellbeing and employment, are present but feature far less frequently. Figure 5 summarises a mode based on eight topics.

Most terms in this model are associated with a single topic (Topic 1) which includes a number of common features and concepts that link most publications. While the terms in Topic 1 tend to be rather general and uninformative, the remaining topics add some useful differential detail.

Figure 6 shows more of the detail of the individual topics, specifically the top six words (defined on the basis of the probability of the word in the topic) in each of the eight topics used to define the model (there are 39 unique words in all, as some words appear in association with more than one topic.)

Topic 1 is the largest and indicates simply that publications commonly refer to data from the UK (often from England) relating to patient health and care. Topic 2 identifies hospital episode and emergency admission statistics as a particular focus. Topic 3 relates to time series data for resource and disease burdens. Topic 4 refers to coronary health in primary care. Topic 5 seems to be something to do with the creation of linked data. Topic 6 relates to orthopaedic surgery and cancer. Topic 7 describes the nature of the data being used: it is commonly retrospective UK

[10] ESRC provides the underlying funding for ADR UK and supports the creation and maintenance of many of the administrative datasets associated with it.
[11] As implemented with the BTM package in R. Briefly, a biterm model uses all pairs of words in a defined span of words as the basis for generating a topic model, and is particularly suited for use with short texts such as publication titles.

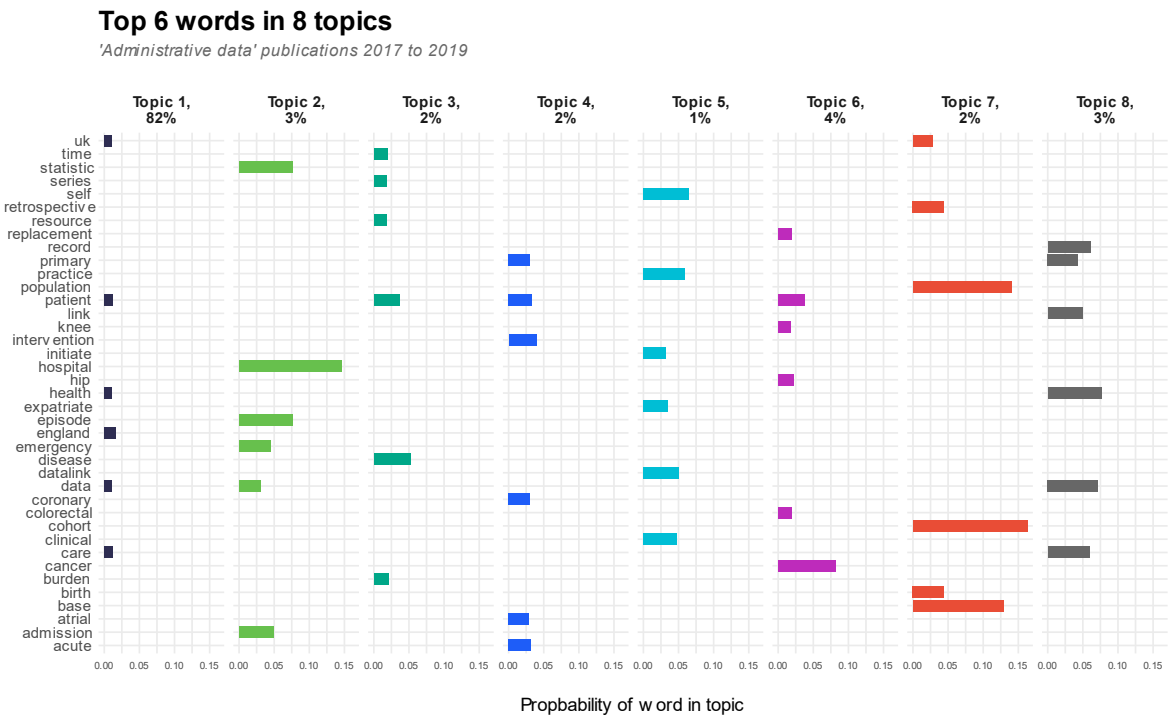

Figure 6: top 6 words in each of 8 topics based on biterm topic model of titles of 'administrative data' publications in 2017 to 2019.

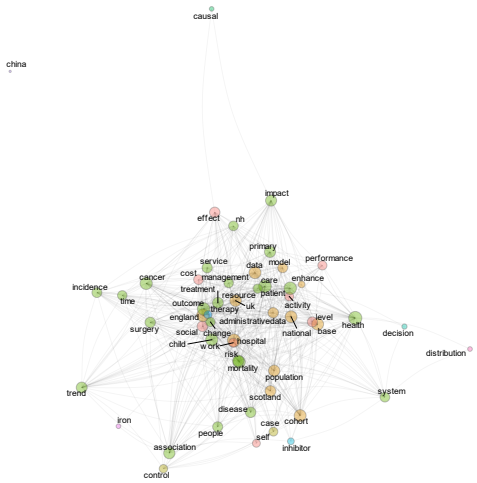

Figure 7:network of links between words comprising biterms in the top 98.5% by total count for titles of UK-affiliated administrative data publications 2017 to 2019

population-based cohort studies. Finally, Topic 8 again refers to the nature of the data as it is primarily about records and linkages[12].

More detailed analysis of the biterms that are the basis for the model further supports the interpretation that the health of the UK's citizens is the main area of interest for these publications. Figure 7 (left) shows the conceptual network of linkages forming the most common biterms.

The core idea, that these publications use data from the UK that relates to hospitals and the health service more broadly, is apparent. The details of some of the specific use-cases, and some other areas of enquiry (such as 'social-work') are also visible.

[12] This interpretation of the results of the model is subjective and, as with any topic model, all topics relate to all publication titles to some extent: a topic model of this kind is a way of identifying things that a publication might be 'about'. Importantly, the composition of the model will reflect varying publication behaviours as well as the actual underlying body of work and so it is not safe to conclude that this dominance in the data reflects real-world dominance.

A similar pattern can be seen in the most common biterms extracted from the set of publication titles (Table 1.)

Word 1	Word 2	Instances
base	population	87
hospital	episode	58
hospital	statistic	58
episode	statistic	54
population	cohort	50
primary	care	48
base	cohort	46
hospital	admission	43
cohort	retrospective	32
hospital	data	30
practice	clinical	27
england	wales	25
care	health	25
people	young	23
electronic	record	22
england	hospital	22
practice	datalink	22
clinical	datalink	22
outcome	patient	21
record	linkage	21

Table 1: most common biterms underpinning the topic model depicted in Figure 5. Biterms are unordered.

In this case, the word ‘base’ is the stem of the word ‘based’, implying that the most common term used to describe research results is that they are ‘population-based’. Hospital episode statistics are very prominent[13], as is the idea of approaching problems with cohort data. The only common biterm that is not obviously related purely to healthcare is ‘young-people’.

It is worth noting that even the most common biterms occur relatively infrequently. The concept of ‘common’ is a relative one.

Conclusions

Administrative data of the kind that ADR UK aims to make available is being used both in the UK and across the globe to support research into a range of issues, primarily those which are related to healthcare. Some other use-cases, relating to broader societal issues, are also present, but not as well-developed.

Some effort is being devoted to the question of how to create and use administrative data effectively, and the researchers engaged are building a base of knowledge relating to this challenge. The results of their efforts are more influential among other researchers than is comparable work in those fields that is not known to use administrative data.

UK-derived data is applied in the contexts to which it is most relevant: the UK itself but also similar wealthier developed nations with advanced healthcare systems. And its use is very much a team sport. Solo efforts are relatively rare, and the field is characterised by medium- to large groups of authors which bring together expertise from multiple organisations. But, as is commonly the case in any area of research, larger organisations have a dominant position, and particularly influential individuals can also be identified.

[13] The nature of the process of generation of biterms means that the underlying phrase ‘hospital episode statistics’ generates three biterms: hospital-episode, episode-statistics and hospital-statistics. It is no surprise that these three terms, derived from the same phrase, appear in the same place in the table and with roughly the same frequency.

Annex 1 – datasets, and data cleaning

The datasets searched for within publication titles and abstracts were developed to reflect those available with the support of ADR UK. The potential for mis-citation, use of name variants (up to three potential variants to be used as search terms were identified for each named data set, see below,) the absence of mentions in either title or abstract and other complicating factors means that the list applied represents a net with a very large mesh and hence likely a biased sample of publications.

Each title was inspected manually to weed out obviously irrelevant entries. However, the rather generic nature of some of the dataset names used (for example ‘Dental Statistics’ and ‘School Census’) means that as well as an unknown number of false negatives, the search results inevitably still return false positives. In these cases the named dataset may not actually have been at the heart of the research.

Weaknesses in the underlying data set are such that this analysis cannot be assumed to be an analysis of the publications resulting from the use of ADR UK datasets.

Dataset name	Variant 1	Variant 2	Variant 3
General Register Office (GRO) Register of Deaths	General Register Office (GRO) Register of Deaths		
Land & Property Services (LPS) Assessment Office (known as Valuation List or Household Valuation Lists)	Land & Property Services (LPS) Assessment Office	Household Valuation Lists	
Agricultural Census in Northern Ireland	Agricultural Census in Northern Ireland		
Higher Education Enrolments	Higher Education Enrolments		
Higher Education Qualifications	Higher Education Qualifications		
Destinations of Leavers from Higher Education	Destinations of Leavers from Higher Education		
Electoral Office Norther Ireland (EONI) Electoral Register (sometimes called electoral roll)	Electoral Office Norther Ireland (EONI) Electoral Register		
GP register (NHAIS)			
Dental Statistics			
Pharmaceutical (EPD)	Enhanced Prescribing Database		
Ophthalmic Statistics			
Benefits Statistics (NI Social Security Benefit database)	NI Social Security Benefit database		
Planning Records Database	Planning Records Database		
School Census			
Schools Leaver's Survey (School-Leavers' Survey)	Schools Leaver's Survey	School-Leavers' Survey	
General Register of Births	General Register of Births		
Environmental Records			
Northern Ireland Longitudinal Study	Northern Ireland Longitudinal Study		
NI Maternity System (The Northern Ireland Maternity System (NIMATS) database)	NI Maternity System	Northern Ireland Maternity System (NIMATS)	NIMATS
GRO infant death	GRO infant death		
NI Enhanced Prescribing Database (EPD)	NI Enhanced Prescribing Database		
Northern Ireland Police Service of Northern Ireland			
Annual District Birth Extract (ADBE)	Annual District Birth Extract	ADBE	
Annual District Death Extract (ADDE)	Annual District Death Extract	ADDE	
Critical Care Dataset	Critical Care Dataset		
Diagnostic and Therapy Services Waiting Times	Diagnostic and Therapy Services Waiting Times		
Emergency Department Data Set (EDDS)	Emergency Department Data Set	EDDS	
National Community Child Health Database (NCCHD)	National Community Child Health Database	NCCHD	

Outpatient Dataset (OPD)		
Outpatient Referral		
Patient Episode Database for Wales (PEDW)	Patient Episode Database for Wales	PEDW
Postponed Admitted Procedures	Postponed Admitted Procedures	
Primary Care GP dataset	Primary Care GP dataset	
Referral to Treatment Times	Referral to Treatment Times	
UK Health Dimensions	UK Health Dimensions	
Welsh Demographic Service (WDS)	Welsh Demographic Service	WDS
Active Adult Survey	Active Adult Survey	
Bowel Screening Wales (BSW)	Bowel Screening Wales	BSW
Breast Test Wales (BTW)	Breast Test Wales	BTW
Cervical Screening Wales (CSW)	Cervical Screening Wales	CSW
Congenital Anomaly Register and Information Service (CARIS)	Congenital Anomaly Register and Information Service	CARIS
Education Attainment		
National Survey for Wales	National Survey for Wales	
Welsh Cancer Intelligence and Surveillance Unit (WCISU)	Welsh Cancer Intelligence and Surveillance Unit	WCISU
Welsh Health Survey	Welsh Health Survey	
Emergency Department Dataset Wales (EDDS) around Accident and Emergency (A&E) attendances	Emergency Department Dataset Wales	
Patient Episode Dataset for Wales (PEDW) around hospital admissions,	Patient Episode Dataset for Wales	PEDW
Welsh Demographics Service (WDS)	Welsh Demographics Service	
NHS Demographic Service Data	NHS Demographic Service Data	
NHS Hospital Admissions		
NHS Hospital Outpatient Data	NHS Hospital Outpatient Data	
Emergency Department Dataset		
All Wales Injury Surveillance Systems (AWISS) dataset	All Wales Injury Surveillance Systems	AWISS
National Community Child Health Dataset (NCCHD)	National Community Child Health Dataset	NCCHD
Pupil Level Annual School Census (PLASC) for attendance records	Pupil Level Annual School Census	PLASC
LLWR (Life Long Learning Wales Record)	Life Long Learning Wales Record	LLWR
WCVA (Wales Council for Voluntary Action	Wales Council for Voluntary Action	

Swansea Child Social Care Data	Swansea Child Social Care Data	
Skills and Employment Survey for Wales	Skills and Employment Survey for Wales	
Supporting People data sets		
Swansea Homelessness Data	Swansea Homelessness Data	
Scottish Index of Multiple Deprivation (SIMD)	Scottish Index of Multiple Deprivation	SIMD
Family Resources Survey 2010/2011	Family Resources Survey	
Poverty and Social Exclusion UK (PSE-UK) 2012	Poverty and Social Exclusion UK	PSE-UK
administrative records from the Scottish Mental Survey 1947	Scottish Mental Survey 1947	
ABS, Annual Business Survey		
AFDI, Annual Inquiry into Foreign Direct Investment	Annual Inquiry into Foreign Direct Investment	
APS (Population), Annual Population Survey		
APS (Purchases) (Forthcoming*)	Annual Purchases Survey	
APS Well-Being	Annual Population Survey: Well-Being	
ARD2, Annual Respondents Database	ARD2	
ARDx, Annual Respondents Database	ARDx	
ASHE, Annual Survey of Hours and Earnings	Annual Survey of Hours and Earnings	
ASHE (pre-release), Annual Survey of Hours and Earnings, provisional 2018 (GB)		
BERD, Business Enterprise Research and Development	Business Enterprise Research and Development	
Births, Births registrations, England and Wales	Births, Births registrations, England and Wales	
BRES, Business Register Employment Survey	Business Register Employment Survey	
BSCI (Discontinued), Business Spending on Capital Items	Business Spending on Capital Items	
BSD, Business Structure Database	Business Structure Database	
BSD Longitudinal, Business Structure Database: Longitudinal	Business Structure Database: Longitudinal	
Capital Stock, Capital Stock		
CEEDR, Research into the Barriers to Take-Up and Use of Business Support	Research into the Barriers to Take-Up and Use of Business Support	CEEDR
Census 1961, Census 1961 Secure Sample (and supplementary files on request)	Census 1961, Census 1961 Secure Sample	
Census 1971, Census 1971 Secure Sample (and supplementary files on request)	Census 1971, Census 1971 Secure Sample	
Census 1981, Census 1981 Secure Sample (and supplementary files on request)	Census 1981, Census 1981 Secure Sample	

hosiCensus 1991 (Forthcoming*), Census 1991 Secure Sample (and supplementary files on request)	hosiCensus	
Census 2011 E&W: Household, Secure Census 2011 England & Wales: Household Sample	Household, Secure Census 2011 England & Wales: Household Sample	
Census 2011 E&W: Individual, Secure Census 2011 England & Wales: Individual Sample	Individual, Secure Census 2011 England & Wales: Individual Sample	
Census 2011 NI: Household, Secure Census 2011 Northern Ireland: Household Sample	Household, Secure Census 2011 Northern Ireland: Household Sample	
Census 2011 NI: Individual, Secure Census 2011 Northern Ireland: Individual Sample	Individual, Secure Census 2011 Northern Ireland: Individual Sample	
Census 2011 OD Safeguarded, Safeguarded Census 2011 Origin / Destination (Flow data)	Census 2011 Origin / Destination	
Census 2011 OD Secure, Secure Census 2011 Origin / Destination (Flow data)	Secure Census 2011 Origin / Destination	
Census 2011 Scotland: Household, Secure Census 2011 Scotland: Household Sample	Secure Census 2011 Scotland: Household Sample	
Census 2011 Scotland: Individual, Secure Census 2011 Scotland: Individual Sample	Secure Census 2011 Scotland: Individual Sample	
CIS, Community Innovation Survey or UKIS: UK Innovation Survey	Community Innovation Survey	UK Innovation Survey
CORE, COntinuous REcording of Lettings and Sales in Social Housing in England	COntinuous REcording of Lettings and Sales in Social Housing in England	
CPI, Consumer Price Index / Retail Price Index		
CSEW, Crime Survey for England and Wales	Crime Survey for England and Wales	
EBS, English Business Survey	English Business Survey	
E-Commerce, E-commerce Survey		
ETB, Effects of Tax and Benefits	Effects of Tax and Benefits	
FALS, Financial Assets and Liabilities Survey	Financial Assets and Liabilities Survey	
GLF, Opinions Survey (formerly General Lifestyle Survey, formerly General Household Survey)	General Lifestyle Survey	General Household Survey
IIA, Investment in Intangible Assets		
IPO, Patents, Designs and Trade Marks from Intellectual Property Office		
ITIS, International Trade in Services		
LBS, London Business Survey	London Business Survey	
LCFS, Living Costs and Food Survey (Expenditure and Food Survey), LCF	Living Costs and Food Survey	Expenditure and Food Survey
LCREE, Low Carbon and Renewable Energy Economy Survey	Low Carbon and Renewable Energy Economy Survey	
LFS Household, Labour Force Survey Household	Labour Force Survey Household	
LFS Longitudinal, Labour Force Survey Longitudinal	Labour Force Survey Longitudinal	
LFS Person, Labour Force Survey Person	Labour Force Survey Person	

LS, Longitudinal Study of England and Wales	Longitudinal Study of England and Wales	
LSBS, Longitudinal Small Business Survey	Longitudinal Small Business Survey	
MES (Forthcoming*), Management Expectations Survey	Management Expectations Survey	
MBS, Monthly Business Survey (merged formerly existed as MIDDS and MPI)		
Mortality, Death Registrations, England and Wales	Mortality, Death Registrations, England and Wales	
MPS, Management Practices Survey	Management Practices Survey	
MQ5, Financial enquiries data		
MWSS, Monthly Wage and Salary Survey (Average Weekly Earnings)	Monthly Wage and Salary Survey	
NEED, National Energy Efficiency data		
NES, New Earnings Survey	New Earnings Survey	
NESS, National Employer Skills Survey	National Employer Skills Survey	
NTS, National Travel Survey	National Travel Survey	
OPSS, Occupational Pension Scheme Survey	Occupational Pension Scheme Survey	OPSS
PPI, Producer Price Index		
PRODCOM, Products of the European Community		
QCES, Quarterly Capital Expenditure Survey	Quarterly Capital Expenditure Survey	QCES
QFI, Quarterly Fuels Inquiry (QFI)	Quarterly Fuels Inquiry	QFI
RPI, Consumer Price Index / Retail Price Index		
SBS, Small Business Survey		
LSBS, Longitudinal Small Business Survey	Longitudinal Small Business Survey	LSBS
SESS, Scottish Employer Skills Survey	Scottish Employer Skills Survey	
SIES, Student Income Expenditure Survey	Student Income Expenditure Survey	SIES
SIPU, Survey of Innovation and Patent Use	Survey of Innovation and Patent Use	SIPU
SoGIS, Survey of International Graduating Students	Survey of International Graduating Students	SoGIS
SRE, Scottish Register of Employment	Scottish Register of Employment	
USoc, Understanding Society	Understanding Society	
Vacancies, Vacancy Survey		
WAS, Wealth and Assets Survey	Wealth and Assets Survey	
WERS, Workplace Employment Relations Survey	Workplace Employment Relations Survey	
Hospital Episode Statistics (HES)	Hospital Episode Statistics	

Opinions and Lifestyle Survey		
Children Looked After Return	Children Looked After Return	
National Pupil Database (NPD)	National Pupil Database	
Veteran Leavers Database (VLD)	Veteran Leavers Database	
Family Justice Data	Family Justice Data	
Affinity Metered Water Sample 2011	Affinity Metered Water Sample 2011	
DECC residential		
British Crime Survey (BCS)	British Crime Survey	
Life Opportunities Survey (LOS)	Life Opportunities Survey	
Opinions Survey (OPN)		
NHS Wales Audiology services dataset	NHS Wales Audiology services dataset	
A&E (Accident & Emergency)		
Air Quality Data		
Social Care Workforce Data (Care Inspectorate Data)	Social Care Workforce Data	Care Inspectorate Data
CHI database		
CHSP-P (Child Health Systems Programme - Pre-School)	CHSP-P	
CHSP-S (Child Health Systems Programme - School)	CHSP-S	
CLAS (Children Looked After Survey)	Children Looked After Survey	
Homelessness Dataset (HL1)		
HAGIS (Healthy Ageing in Scotland) survey	Healthy Ageing in Scotland	
NASCAR (Northeast and Aberdeen Scottish Cancer and Residence)	Northeast and Aberdeen Scottish Cancer and Residence	
NCA (National Crime Agency - Seized Packages)	National Crime Agency - Seized Packages	
NHS24		
NHSCR (NHS Central Registry)		
NRS Census 2001	NRS Census 2001	
NRS Census 2011	NRS Census 2011	
NRS Deaths	NRS Deaths	
PIS (Prescribing Information System)	Prescribing Information System	
School & Pupil Census		
School Leaver Destinations		

SCI-Diabetes (Diabetes Register)	SCI-Diabetes
SAS (Scottish Ambulance Service Attendances)	Scottish Ambulance Service Attendances
SDMD (Scottish Drug Misuse Database)	Scottish Drug Misuse Database
SLS (Scottish Longitudinal Survey)	Scottish Longitudinal Survey
SMR00 (Scottish Morbidity Record - Outpatient Appointments & Attendances)	SMR00
SMR01 (Scottish Morbidity Record - General Acute Inpatient & Day Case)	SMR01
SMR02 (Scottish Morbidity Record - Maternity Inpatient & Day Case)	MSR02
SMR04 (Scottish Morbidity Record - Mental Health Inpatient & Day Case)	SMR04
SMR06 (Scottish Morbidity Record - Scottish Cancer Registry)	SMR06
SMS47 (Scottish Mental Survey 1947)	SMS47
SSCA (Scottish Stroke Care Audit)	Scottish Stroke Care Audit
Social Care Survey	
UCD (Unscheduled Care Datamart)	Unscheduled Care Datamart
OOH (GP Out of Hours)	
MOD Service Leavers Database	MOD Service Leavers Database
SCRA (Scottish Children's Reporters Administration)	Scottish Children's Reporters Administration

Annex 2 – Dimensions data, cleaning and reproducibility of this analysis

Using the list of datasets in Annex 1 it should be possible for anyone to reproduce, approximately at least, the raw data used in this analysis by use of a suitable search query in Dimensions. Free access to the Dimensions system for scientometric studies can be requested at https://www.dimensions.ai/scientometric-research/ (although as of September 2020 this has been limited slightly due to Covid.)

UKRI has paid access to a private instance of Dimensions which in principle would allow more advanced querying of the full text of publications for the search terms. But the aim here is to produce a more open piece of work, even at the expense of accuracy, and so only the publicly-available simple querying (not using the Dimensions API) has been used. The imprecise nature of the concept of 'the use of administrative data' and limitations of the querying process, however advanced, mean that it would never be possible to produce a perfect picture of all outputs that use administrative data anyway.

Data are used as received and no attempt was made to fill in missing data (for example, author affiliations.) Data cleaning had two main steps: first, the removal of irrelevant publications and second the disambiguation of authors. Both steps are subjective, meaning that the precise details of the analysis probably could not be reproduced exactly (even by the author) if started over from scratch. However, all cleaning undertaken is documented in code, as are all the other numerous data-related choices made, including those related to the topic modelling. Topic models are inherently unreproducible in their details and so the model described should, as with everything else, be taken as indicative only. All analyses were conducted in R.