Open data in the health sector

Users, stories, products and recommendations

Giuseppe Sollazzo and David Miller

Published February 2017

OPEN HEALTH CARE UK

Table of Contents

Executive Summary

NHS England have commissioned Open Health Care to gather evidence from the community of health open data users to help understand the needs of people using data outside NHS England. This report is intended to provide clear recommendations and evidence of user need to inform NHS England's open data policy and programme development.

This report, written by Giuseppe Sollazzo in collaboration with David Miller of Open Health Care, outlines what we have learned from users of health open data in the health sector and beyond.

Following our interviews, discussions and subsequent analysis, we present two specific recommendations, as well as some general principles for data publishing with concrete examples.

Recommendation 1

Work towards a single authoritative online point of access for datasets in the health sector, and proceed with a programme of reducing duplication in data publishing. This online service should include both direct access to open data, and a clear process to access shared data.

Recommendation 2

Establish an expert and reachable data support team with a remit covering the entire health system, not limited to individual institutions. This team should be encouraged to build strong links with the wider data ecosystem and lead a health data user group.

We discuss these recommendations and their rationale in further detail below.

Introduction

The phrase open data has been part of the transparency activists' vocabulary for over ten years, and governments around the world have been seen to engage enthusiastically with the open data policy agenda. The Open Government Partnership has demanded and agreed clear commitments on open data from participating countries and embedded them into their National Action Plans.

In the United Kingdom, Government departments and agencies now routinely publish open data and often issue press releases to advertise their achievements. The central Government data portal data.gov.uk has lived through several phases, and is showing renewed interest while undergoing a user research review.

NHS England have been heavily involved with the open data agenda in the health system, both releasing data and collaborating with other health agencies in order to support their data programmes. The UK health sector has seen increasing numbers of data releases from the likes of Public Health England, the Care Quality Commission, NHS Digital (formerly the Health and Social Care Information Centre), and other Department of Health arm's-length bodies. These programmes and collaborations were also encouraged between 2012 and 2015 by the Health and Social Care Transparency Panel, an advisory board of the Department of Health chaired by the then Undersecretary of State Dr Dan Poulter MP.

More recently, NHS England launched the NHS England Data Catalogue, an online data portal containing open datasets routinely used by NHS England.

What do we mean when we talk about open data?

There are multiple definitions of "open data", some of which are relatively strict. TheOpen Definition from Open Knowledge for example, is technical, specific and subject to versioning, but may be summarised as:

Open data and content can be freely used, modified and shared by anyone forany purpose.

The extended version mandates that to be deemed "open", a licence needs to grant an extensive set of required permissions and sets out acceptable extra conditions that may be imposed.

In this report we have chosen not to limit our discussion to open data. We discovered early in our research that in practice, analysis, products and services that use health data move freely across the Data Spectrum as dictated by their individual needs.

Much health data is not published as open data, frequently because of the sensitive and personal nature of information about our health. Users of health data will often combine open data with data that is only available to them as the holder (closed data), or data that they are able to access because they meet certain criteria (shared data).

Any programme or policy effort that ignores the wider context in which data is used, addressing open data exclusively will likely fail to achieve its goals when it meets the messy, pragmatic compromises that data use in the real world consists of.

Accordingly, while we have continued to focus mainly on open data, we have included in this report the results of our research pertaining to all health data, open, shared and closed, published by any public sector body.

Approach

We contacted a wide network of professionals working with health data including developers, general practitioners, academics, researchers, entrepreneurs, private sector company leaders, and government officials. We spoke with over 20 people, and conducted 11 formal structured interviews. Summaries of these interviews are included in this report. The remaining contacts provided ideas, introductions, or useful information via e-mail.

The people we interviewed belong to four broad areas

- Users of data for research purposes, to experiment for academic or professional development reasons
- Users from businesses who offer data related products or services
- Users who create evidence to support their line of business
- Users who are medical professionals and use data to audit their activity

All participants were asked to respond to three wide-ranging questions:

1. What do you/your colleagues use health open data for? (or: are you aware of interesting uses of health open data?)
2. Do you find any issues in the data regarding quality, access, process, or any other problem?
3. Do you have any recommendations for the data publishers?

These questions were used to kickstart the conversations. Although we took notes, these are not included in the report, where we include a structured, but narrative summary of the findings.

In addition to the one-to-one e-mails and interviews, a general survey was launched via Twitter and disseminated through several data and health practitioners' social networks. The survey asked roughly the same three questions. Despite a large number of shares on social media, the survey collected few responses. We factored these into the general analysis.

Interviews

In this section, we give a brief summary of each interview, following a generic structure: biographical notes on the interviewee, main datasets they mentioned, example of data uses, issues reported, and recommendations that emerged from the conversation. In some instances we discussed general issues regarding health Open Data, due to the nature of the interviewee's work, and we report a summary that does not follow the general structured scheme.

Anna Powell-Smith, Academic Software Developer at EBM Data Lab

"We use Prescribing Data to produce datasets for researchers"

At the time of the interview, Anna Powell-Smith was tech lead at the Evidence Based Medicine (EBM) Data Lab at the University of Oxford. The lab is led by Ben Goldacre. EBM Data Lab *build working, useful products to help academics and doctors. [They] campaign for better data in healthcare.*

Datasets

Anna lists a number of datasets which contribute to this:

- Practice List Size and GP Count for each practice, from the NHS Business Services Authority
- Clinical Commissioning Group boundaries, from the NHS England Resources website (and previously from the ONS Open Geography Portal)
- GP Practice Dispensing Status (now vanished and replaced by the Dispensing Practice Name and address, from NHS Prescription Services)

Data Uses

The lab has recently focused on drug prescribing and clinical trials data. They aggregate the prescribing data from NHS Digital into an openly accessible database called OpenPrescribing, which in turn they use to drive their research, building tools for GPs and medicines managers. Most of the outcomes of their work are bespoke datasets exports, and data analysis tools which are used within the health system.

Their product is derived data, which has enabled interesting reuses of it. The OpenPrescribing dataset is especially popular with researchers. Several academic publications have appeared that use this database. For example, this abstract poster from Rheumatology.

Issues

Anna reports several issues encountered in their work:

- The data from BSA data is attached to an Open Government Licence, but is weirdly protected by a captcha/login
- The practice list size data from BSA has different granularity from the prescribing data, which makes it difficult to link different datasets
- Weightings per disease are not released by BSA
- Often there isn't a single place where to look for data, there are often multiple sources for health related data, and all apparently canonical
- Slices of data are not under OGL
- BSA access is split between NHS and not-NHS users, because of privacy issues; however, there are 1.2 million people working in the NHS, so this seems to be not a serious issue (or one that could be dealt with in a better way).

Recommendations

Anna made some specific suggestions:

- Different formats for different datasets; big datasets should be released for example using BigQuery exports rather than CSV or XLS files

- There is a need for support and engagement on the NHS side; this should be similar to what happens in the ONS, where each dataset is connected to a named officer who can support and advice; an SLA to answer question (48 hours). NHS Digital are generally good at this, while BSA take 28 days to respond: a single unit could help.
- Lists like GP surgeries and CCG boundaries should be managed “as if they were registers”, with a clear accountability chain
- Consistent codes need to be used across datasets and cross-checked
- Public pages with good SEO where the datasets are published, not behind captchas like the BSA data
- Do not restrict dataset access to internal NHS users without good reason (a load of BSA data is only available to NHS IPs)
- Decent README and documentation with dataset
- Use standard open licences
- Someone at NHS Digital whose full-time job it is to help people navigate datasets, and who can help the organization fixing issues found in the datasets.

Callum Tanner, Public Health Analyst at Isle of Wight Council

“Health open data helped me conduct the Isle of Wight Joint Strategic Needs Assessment”

Callum Tanner is a Data Analyst for the Isle of Wight Council, where the local authority is responsible for all of the Public Health initiatives. Similarly to other Local Authorities, the Isle of Wight has started using data analysis units as a way to inform their operations. In March 2015 the Isle of Wight was selected as one of the first 29 Vanguard sites in Britain which released an additional £200 million worth of NHS England funding to integrate home care, mental health and community nursing, GP services and hospitals for the first time since 1948. Callum produces data analyses for an integrated health and social care system on the Isle of Wight.

Datasets

Callum lists a series of datasets used in several projects:

- Health Impact of Physical Inactivity (HIPI) Tool from the South West Public Health Observatory
- Active People Survey from Sports England
- National Child Measurement Programme from Fingertips/NHS Digital
- Breastfeeding Statistics for initiation and 6-8 weeks from Data.Gov.UK
- Health Visitor Data from NHS England and the National Child and Maternal Health Intelligence Network (PHE)
- NHS Dental Practices from digital.nhs.uk

Data Uses

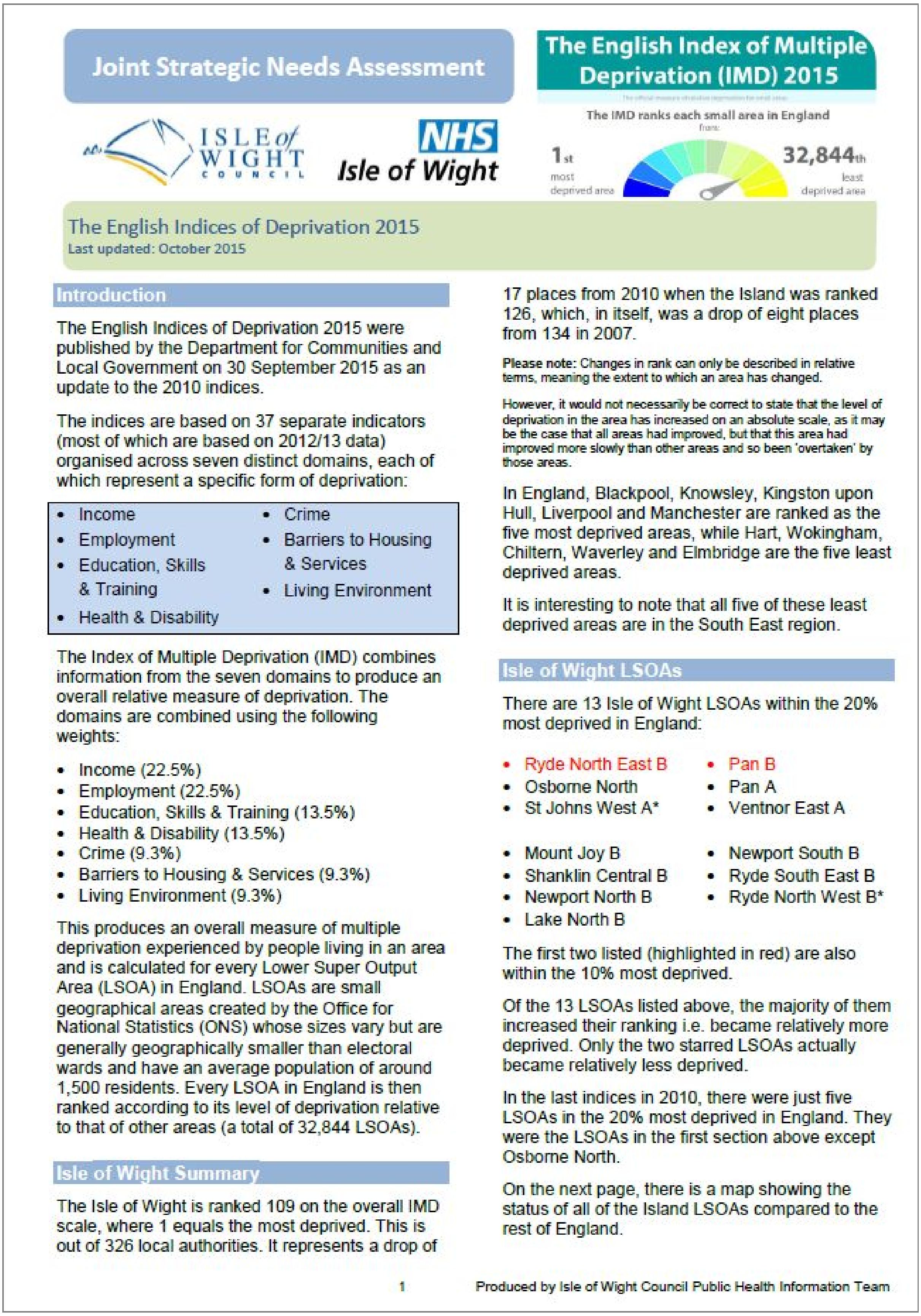
Callum used a variety of datasets to conduct the Isle of Wight Joint Strategic Needs Assessment (JSNA). The results of this JSNA are published .

For example, the Physical Activity aspects of the JSNA were conducted using the Health Impact of Physical Inactivity (HIPI) Tool, released by the South West Public Health Observatory and the Active People Survey from Sports England.

An important part of the JSNA was the Obesity assessment, for which the National Child Measurement Programme Data was used at both Ward and Local Authority level. This was accessed through the Fingertips Tool and the NHS Digital Website.

OpenPrescribing data was used it to see how many IUD coils had been provided in the area.

A more simple use that comes to mind is a list of NHS dental practices from digital.nhs.uk. This was used by Callum to create a combo box limited to values from a lookup table in an Access database. *“We have a lot of examples of small uses of health data like this one”*, says Callum.



Joint Strategic Needs Assessment

ISLE of WIGHT COUNCIL

NHS Isle of Wight

The English Index of Multiple Deprivation (IMD) 2015

The official measure of relative deprivation for small areas

The IMD ranks each small area in England from:

1st most deprived area

32,844th least deprived area

The English Indices of Deprivation 2015

Last updated: October 2015

Introduction

The English Indices of Deprivation 2015 were published by the Department for Communities and Local Government on 30 September 2015 as an update to the 2010 indices.

The indices are based on 37 separate indicators (most of which are based on 2012/13 data) organised across seven distinct domains, each of which represent a specific form of deprivation:

- Income
- Employment
- Education, Skills & Training
- Health & Disability
- Crime
- Barriers to Housing & Services
- Living Environment

The Index of Multiple Deprivation (IMD) combines information from the seven domains to produce an overall relative measure of deprivation. The domains are combined using the following weights:

- Income (22.5%)
- Employment (22.5%)
- Education, Skills & Training (13.5%)
- Health & Disability (13.5%)
- Crime (9.3%)
- Barriers to Housing & Services (9.3%)
- Living Environment (9.3%)

This produces an overall measure of multiple deprivation experienced by people living in an area and is calculated for every Lower Super Output Area (LSOA) in England. LSOAs are small geographical areas created by the Office for National Statistics (ONS) whose sizes vary but are generally geographically smaller than electoral wards and have an average population of around 1,500 residents. Every LSOA in England is then ranked according to its level of deprivation relative to that of other areas (a total of 32,844 LSOAs).

Isle of Wight Summary

The Isle of Wight is ranked 109 on the overall IMD scale, where 1 equals the most deprived. This is out of 326 local authorities. It represents a drop of 17 places from 2010 when the Island was ranked 126, which, in itself, was a drop of eight places from 134 in 2007.

Please note: Changes in rank can only be described in relative terms, meaning the extent to which an area has changed.

However, it would not necessarily be correct to state that the level of deprivation in the area has increased on an absolute scale, as it may be the case that all areas had improved, but that this area had improved more slowly than other areas and so been 'overtaken' by those areas.

In England, Blackpool, Knowsley, Kingston upon Hull, Liverpool and Manchester are ranked as the five most deprived areas, while Hart, Wokingham, Chiltern, Waverley and Elmbridge are the five least deprived areas.

It is interesting to note that all five of these least deprived areas are in the South East region.

Isle of Wight LSOAs

There are 13 Isle of Wight LSOAs within the 20% most deprived in England:

- Ryde North East B
- Osborne North
- St Johns West A*
- Pan B
- Pan A
- Ventnor East A

- Mount Joy B
- Shanklin Central B
- Newport North B
- Lake North B
- Newport South B
- Ryde South East B
- Ryde North West B*

The first two listed (highlighted in red) are also within the 10% most deprived.

Of the 13 LSOAs listed above, the majority of them increased their ranking i.e. became relatively more deprived. Only the two starred LSOAs actually became relatively less deprived.

In the last indices in 2010, there were just five LSOAs in the 20% most deprived in England. They were the LSOAs in the first section above except Osborne North.

On the next page, there is a map showing the status of all of the Island LSOAs compared to the rest of England.

Issues

Callum reports situations in which the licensing is not clear and a lot of confusion as to where the data should be official. Callum also suggests aggregate datasets from sources like OpenPrescribing are a powerful tool because of information governance issues:

Health Visitor data is a commissioned service by Public Health England. It is now updated on the CHIMAT website but it is reported that PHE are moving all content to the gov.uk domain. Maybe this has happened already but the data is updated on CHIMAT at present. All of this data will form part of the Maternity Children's Data Set (MCDS) but this is still in its infancy and the current processes could continue up until 2017/18 which is probably to allow time to improve data standards across all the datasets of MCDS. Trying to get my head around where all the information is has been quite frustrating.

Given some of the perceived and real IG barriers across NHS, GP's and the Council which can sometimes be a hindrance, regularly updated intermediate websites like openprescribing.net, or fingertips can be extremely useful.

For our commissioned services it is easier to access rich record level datasets as we have data sharing processes and agreements in place in order to monitor KPIs, but for services which are not directly managed by Public Health, open data is often our best source of information.

Recommendations

Two themes emerged from our conversation with Callum: firstly a strong desire to see a single point of access for all the UK health data; secondly, that every dataset should have its licensing clearly marked.

Edafe Onerhime, Open Data Services Coop

"We use open data to Connect people who are building innovative health services"

Edafe Onerhime is a data expert with a Masters in Business Intelligence. During her studies she tried to answer the question "How good is health open data?" She subsequently worked on freelance data related projects and runs data dives, events, data literacy workshops, and data for good events through ACT Collective (a collaborative practice focused on the charity sector). She is an ODI registered trainer and currently works with Open Data Services, a digital co-operative, working on projects for social good.

Although Edafe doesn't directly make use of health open data, she has worked through Open Data Services on a number of grant-based projects. In this respect, Edafe insists that the most interesting aspect is the release of Contracting Data in the Health Sector using the Open Contracting Data Standard, which is being widely adopted by the civil service at large.

The major goal in this would be the ability to "follow the money" through standards and open data and offer health services and public authorities the opportunity to analyse grants given to and from hospitals, keep accurate accounts of human services (e.g. volunteers numbers and skills, etc), and acquire the ability to link contracts to grants, and grants to human services.

Edafe is excited to develop a discussion around canonical registers.

However, Edafe is also adamant that her work increasingly suggests that for the data to be useful, it should be first used by health professionals; this however requires bi-directional workflows and data capability building.

As far as Open Data Services are concerned, there are several areas of interest in health open datasets to be released. Edafe provided an overview of the types of datasets required based on projects she has been working on.

Open Contracting is an initiative advocating the adoption of open standards and data releases in all phases of Government procurement. In this respect, there are several areas of interest for data releases, for example:

- Hospital and other health construction and infrastructure
- Property Management especially private finance initiative (PFI)
- Other public–private partnerships (PPPs)
- NHS Procurement including medication
- NHS Budget & Spend

360Giving is a project that supports organisations to publish their grants data in an open, standardised way and helps people to understand and use the data in order to support decision-making and learning across the charitable giving sector. Regarding health data, the project would benefit from data about grants to and from NHS organisations, and details of beneficiaries.

I know that there is a grants agreement for voluntary organisations but I cannot easily pinpoint beneficiaries.

A similar initiative, Open Referral, aims to develop standards and platforms to share community resources information, where data such as human services in health care and volunteering would be highly beneficial.

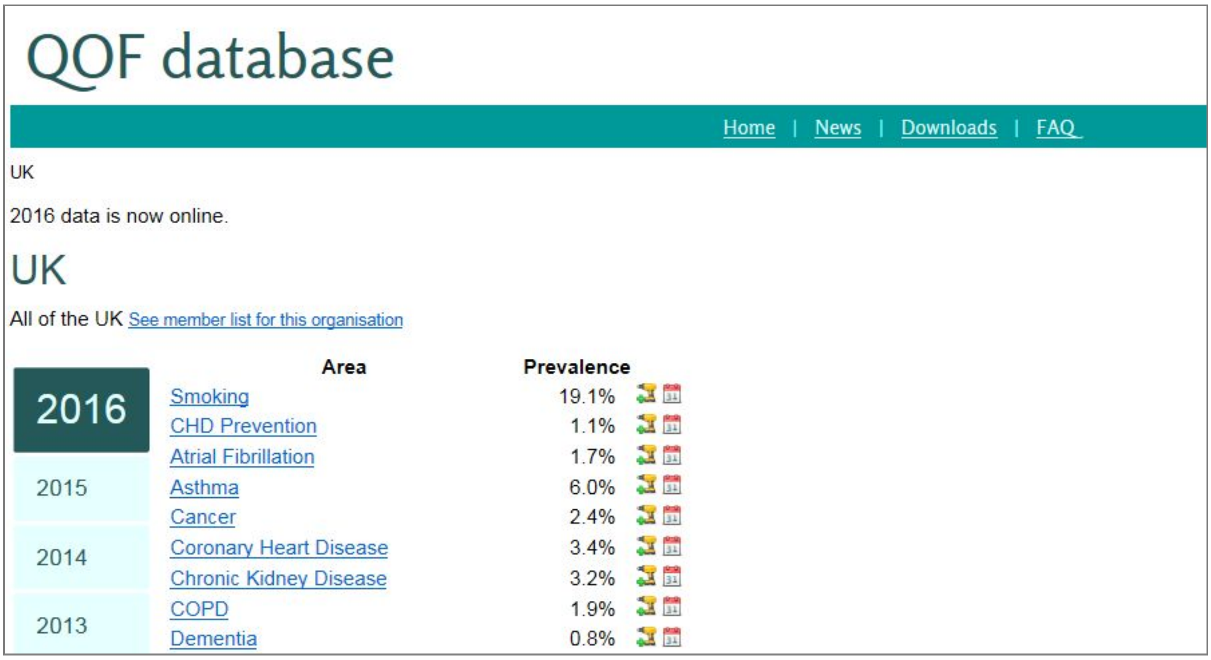
The Joined up Data Alliance is a project run by Omidyar Network that supports sustainable development by sharing data about organisations. Some pressing issues they find are:

- How can an NHS organisation be identified?
- There are multiple places to find a hospital or other organisation.
- There is no official registrar to focus on defining the entities and keeping the data up-to-date.

Gavin Jamie, General Practitioner

"I use QOF data to compare prevalence between my patients and other areas of the country, and to release aggregate calculations for the constituent countries of the United Kingdom"

Gavin Jamie is a General Practitioner. His practice is in Swindon and he also has a MSc in Healthcare Informatics from the University of Bath. Gavin also runs a service that aggregates Quality and Outcomes Framework (QOF) data, where data going back to 2005 can be found under a Creative Commons Licence, presented in several formats including CSV and Access DB, the latter proving particularly popular with other GPs.



QOF database

Home | News | Downloads | FAQ

UK

2016 data is now online.

UK

All of the UK See member list for this organisation

	Area	Prevalence
2016	Smoking	19.1%
	CHD Prevention	1.1%
	Atrial Fibrillation	1.7%
2015	Asthma	6.0%
	Cancer	2.4%
2014	Coronary Heart Disease	3.4%
	Chronic Kidney Disease	3.2%
2013	COPD	1.9%
	Dementia	0.8%

Datasets

Gavin uses a variety of sources both for his General Practice activities and his QOF aggregation service:

- Own statistical data from his GP Practice
- QOF from NHS Digital
- Prescription data from NHS Digital

It should be noted that Gavin is a long time data user. He has been collating and re-sharing QOF data on his website since 2004, before the age of open data. *"I used to do 200 FOI requests a year to get data from PCTs"*, Gavin recalls. He notices how things have changed for the better: *"Once I was even contacted a month ahead of publication and asked which formats would be best for me"*, Gavin adds.

Data Uses

The most important way Gavin uses this data is for auditing his practice. *"We ask questions"*, he says, "such as: 'Are we missing cases of certain conditions?' or 'Are we prescribing very differently from other surgeries?'". Gavin describes a number of population health studies done using data. For example, he analysed data on blood pressure at his GP surgery in Swindon, which has a total of about 10,000 patients. Of which, about 1,000 suffer from high blood pressure, which is near the national average.

However, Swindon is demographically different from rest of Wiltshire, with a considerably higher migrant population and fewer people living in rural areas. By using QOF, Gavin was able to verify that the most similar prevalence for high blood pressure were in Milton Keynes and Bristol, raising interesting questions around the management of public health campaigns.

On another instance, they found out they had higher prescription rates of an antibiotic compared to the national average.

We had an issue with a potential payment under a Local Enhanced Service - essentially a mini contract with incentives for hitting specific target. The target in question was about prescribing certain types of antibiotics, based on a national target. The denominator is all prescriptions for antibiotics and my practice has tended to be above average on that ratio. We were able to use OpenPrescribing data to show that we actually prescribed fewer of these antibiotics than almost every other practice in Swindon. The problem was that we did not prescribe enough other antibiotics in the denominator to bring down the ratio. The CCG paid out.

Issues

Gavin presented a number of issues for the datasets he uses, most of which are due to the level of aggregation. He complains that at the granularity the data is made open, it is often difficult to interpret.

QOF gives totals for conditions/prescriptions, without the ability to link. For example: if a patient has been given an antibiotic and a flu jab, they will appear under two totals. There is no way to see which patients received both. This seems to be a recurrent problem, another example of which is national diabetes data: at the current level of aggregation, it is definitely good for the press to create news about it, but not good enough to improve outcomes for patients, especially for a GP.

A more technical complaint is the lack of properly linked data; for example there is a lot of data about manpower in practices and payments to practices, but these are in different files, split between multiple files. Gavin suggests that it would be good for GPs to have access to tools that assist them with practical tasks related to analysing the data.

Recommendations

I only have one recommendation: give us data we don't have.

An example is hospital admissions by GP practice. At the moment this is only available from the Hospital Episodes Statistics dataset (HES) which is incredibly hard to access for GPs. Gavin reports that such data was routinely accessed by GPs ten years ago, but that data governance has made this increasingly difficult.

Jamie Whyte, Trafford Council Innovation and Intelligence Lab Manager

"We used open data to increase cervical cancer screening across the Borough"

Jamie Whyte works for the Trafford Council in Greater Manchester, where he set up an Innovation and Intelligence Lab with the help of the Release of Data Fund administered by the Open Data User Group and Cabinet Office. The Lab is a multi-disciplinary team of data specialists developing innovative ideas, often based on data, to reinvent Council-run services. As their website says:

We use all sorts of data - open, closed, big and linked - and turn it into intelligence that is used to (re)design services, understand demand and inform citizens. At the same time we are testing and exploring new technologies to improve the way that we do things.

Jamie is the Head of the Lab and a co-lead on open data for LocalGov Digital.

Datasets

- Fingertips
- Local health profiles
- Prescribing data
- Non-open data from GP Practices
- North West Ambulance Response Data

Data Uses

Jamie reports that the aggregation level of open data is often an obstacle. There is a lot of data that could be released with an Open Licence at a lower level.

We don’t do a huge amount with health open data - partly because it mostly doesn’t seem to go down to small area levels.

Small area data would be really useful in targeting local health campaigns. For example in Trafford, we have an issue with high rates of hospital admissions for unintentional and deliberate injuries of 0-4 year olds. If that data were open at small area level, it would be much easier to share with partners, and encourage third sector organisations to apply for money and design projects to help tackle the problem. This would allow us all to target interventions better. As it is, the data is held in databases that are not open (probably rightly so), but they could output data for the country showing this sorts of data at small area - for example MSOA / LSOA / Ward.

Fingertips is used regularly to produce maps. An example is data about smoking, analysed using Fingertips to identify the location of smoking clinics. This data is useful because it gives data at a GP Practice level, which is useful to map events and show disparity within the borough. They also use local health profiles, which provide ward level data.

Prescribing data was used to look for mental health issues by analysing ratios of antidepressant prescriptions in order to inform social workers activities. In this instance the team really struggled to process the data, suggesting they would need a high performance computing platform in order to do this in a reasonable time.

Cervical cancer screening received some attention too, although this required putting together health data with non-health data, for instance population, demographics, languages. They used the data to identify areas with the lowest cervical cancer screening rates, and then profiled those areas in question using demographic data. To print leaflets, they used data about languages spoken, ethnicities religion, housing types, and housing providers.

Some closed data from GP Practices was also used, for example postcodes of women who were invited for cervical cancer screening, but didn’t attend. They mapped these to give commissioners a group of streets to focus on, and they went into the community knowing what to expect. It was a successful project achieving a big increase in screening rates including women who had problems identified at the screening and treated, who would have not otherwise been screened using the previous approach.

A similar successful project managed to increase the number of defibrillators in the borough. Using data to profile areas - demographics, cardiovascular disease rates, physical activity, and ambulance call-outs allowed the council to produce a priority list of areas to put defibrillators. Businesses were contacted from the Mayor’s Office to ask for donations to install a defibrillator close by. These engagement activities, raising awareness, together with other charities’ work led to a massive increase in the availability of defibrillators.

Issues

Jamie reports difficulties around the licensing on Fingertips; it is not clearly stated, and as a consequence there is often little confidence that the data is open or what licensing is attached to it.

Each dataset seems to have separate usage rules. In one case they found a request to add disclaimer stating: “This material is Crown Copyright but may be reproduced without formal permission or charge for personal or in-house use and should be acknowledged as ‘© Crown Copyright, source: Public Health England 2015’.”

Local Health data has licensing issues.

We can’t tell whether it’s open data, and it can be tricky to use this. An API or similar would be lovely.

Jamie sent us an example of two different levels of usability: "Look at these two websites. This looks like it's shaping up to be pretty good. But this is horrible."

One common complaint about health open data is that it doesn't go down enough to a small area level. Enabling operations for councils often requires street-level data, but having them on a council ward level would be good enough in most cases.

Recommendations

Jamie offered several recommendations in three areas: firstly clear and explicit licensing of each dataset, so that the local government officers can use the data with confidence; secondly, the ability to access data at a lower level of aggregation whenever possible; thirdly, a technical recommendation around data access via API, especially for data that changes often, or identifying codes that need to be verified. This could be useful in all those cases where identifier can be short-lived. For example, given the BNF code for a certain medicine, an API method could allow the user to verify whether that code is still valid, or if that drug is no longer being dispensed via the NHS; or a similar example could apply to the GP surgeries list: an API call would let the user know the status of a certain practice, relaying the fact that it is no longer operating. This kind of API would enable users to execute quick verifications of values in their applications.

Judy Aldred, CEO, SSentif

"We use data to power dashboards that give our clients insight on their operations"

Judy is the founder and CEO of SSentif, a company providing intelligence and data analysis services to NHS organisations and Local Government. They offer an online benchmarking tool which provides over a million indicators, targets and alerts, visualizations, outlier detections, and other similar features. Their major clients are NHS Trusts, CCGs and Local Authorities.

Datasets

SSentif is powered by a variety of data sources. They are not limited to open data, but use any data they can access.

We use anything available, any available dataset is good for us; if there's a process to access it, it is open enough for us.

This has previously included Hospital Episodes Statistics dataset, although access to it has become more strict in recent years.

Judy lists a "Top 4" of their most used datasets:

- Prescribing data for CCGs
- Hospital mortality
- Adult social care data
- Estates Returns Information Collection (ERIC) from NHS Digital for fire alarm data

Data Uses

SSentif's sell a dashboard of data analytics and visualizations to several authorities and agencies in the country. The data is grouped by areas of competence, and each dataset is used to create an indicator and a relevant visualization.

Their product is used for benchmarking and bringing data together: health, social care, planning, employment, education, transport. Indicators about a wealth of factors are available and visualizations are ready too. SSentif offers ready-made dashboards, plus the ability to build personal and custom ones.

Ssentif clean the data (missing, quality issues, etc). If problems are identified in the data, they get back to the publisher and ask for corrections or clarifications. A relationship has been built over the years, so they generally avoid using FOI.

Judy reports an increase in the willingness to use data to take action by authorities and NHS Trusts, compared with the situation 7 or 8 years ago.

Issues

There are several feedback issues with some of the authorities. Judy reports good management of requests from NHS Digltal, but attributes prompt replies to her years of engaging with the institution rather than a general breakthrough in engagement levels.

There isn't enough community health data, which is requested a lot by their customers; on the other hand, there are many sources of mental health data, but often in formats that is difficult to use, lacking standardised publication formats, and often published across many files. Sometimes this is caused by the frequent changes in the mental health realm, causing structural changes to the data that make it difficult to use the data longitudinally.

Judy uses the adjective "painful" when she talks about organisation-level data. She would like to see some standard mechanism to receive information about what is new in terms of organizations. For example to be alerted when hospitals merge, with guidance on how to interpret historical data. Metadata is generally ok, although dataset descriptions can always be improved.

Psychological professional data such as IAPT (Improving Access to Psychological Therapies) is also difficult to use, as it is at a trust level, disseminated in multiple files.

Recommendations

Judy's wish list starts with fast replies, even in absence of a named contact; she reports that NHS Digital used to be excellent at this, responding to queries in 24-48 hours.

Another important point is that of datasets being split in too many files: even when data is generated at a Trust level, it should be made available as a national dataset.

Mark Barrett, Co-Founder of DataFlock

"We use open data to offer our customers data analysis tools"

Mark Barrett is co-founder of DataFlock. Previously of HebeWorks, he founded the then LeedsDataMill (now DataMillNorth), which received some funding from the Release of Data fund. Previously he reached open data fame thanks to his GP ratings app. Mark has been recognised as one of the top 50 Innovators in Healthcare by the Health Service Journal, selected as one of the Top 50 "New Radicals" by Nesta.

DataFlock is a data science consultancy, using open data to build tools and reports for a variety of customers. At the moment they are working with CCGs. DataFlock's approach is, in Mark's words, to "seek any data set available out there", rather than advocating for any specific data release. They use the NHS England Data Catalogue, Data.Gov.Uk, NHS Digital, and the Indicator Portal which Mark helped develop when he worked for NHS Digital.

Datasets

- "NHS RightCare – Commissioning for value" packs, published by NHS England
- Monthly Prescription Data, published by NHS Digital
- Annual Quality Outcomes Framework data, published by NHS Digital
- The GP Patient survey published by NHS England
- LSOA boundary shapefiles, published by Office for National Statistics.
- National Census data, published by Office of National Statistics
- Childhood obesity data from the National Child Measurement program
- Adult obesity data from NHS Digital – QOF
- GCSE results data from Department of Education
- LSOA and Ward, and population estimates from ONS
- Takeaway restaurants ratings from Food Standards Agency

Data Uses

DataFlock were recently commissioned by Bradford Districts CCG to investigate respiratory conditions across the three CCGs in the area as part of the "Bradford NHS Open for Innovation" event by Digital Catapult, Medipex at DHEZ. They used a variety of datasets, including BNF and prescribing, census and shapefiles, to analyseRespiratory Patterns in Bradford's CCGs.

Mark and his colleagues looked at the "Commissioning for value" data packs to detect the three lowest performing CCGs for each sub-topic (e.g. Asthma, COPD). They then used 6 years of monthly prescription data related to respiratory conditions, and selected practices within the Bradford CCGs to detect both seasonal trends and practices that are outliers. This was combined with GP Patient Survey data and deprivation levels from the Index of Multiple Deprivation to detect the potential for issues in the way the practices deal with respiratory diseases, and used census data about languages to design interventions intended to reduce A&E Attendance by targeted campaigns. DataFlock also ran an analysis of Child Obesity Levels by exposing that only 1.8% of council wards in England are reducing childhood obesity.

Issues

Mark finds that search is badly implemented in data portals and prefers to use Google. He finds, however, that things are improving with the advent of the NHS England Data Catalogue, which marks an improvement from the time when datasets where only available on the NHS Digital website.

Recommendations

Mark recommends moving everything under a single point of access for data coming from different agencies.

There seems to be a preference for CSV rather than, as Mark puts it, "exotic formats" or linked data.

We would simply like to avoid things that waste a lot of time, for example there is no reason to release Excel files, especially with multiple sheets, and drop down items that manipulate the data and create graphs. We spend a lot of time stripping out this functionality to get to the raw data - a basic CSV file would save us a lot of time as we manipulate the data using our own methods, often for different purposes.

In terms of support and engagement, from talking to Mark it emerged that named contacts or response unit per dataset would help, as well as offering a way for collective engagement ("a Slack channel, for example") with all data users.

Mark also suggests showing more case studies on the catalogue: many data users would appreciate exposure, especially if promoted further via social media.

Marcus Baw, Locum General Practitioner

"I enjoy using medical data for hackdays that show the potential for innovation"

Marcus Baw is a locum GP in the North West of England, who combines his work as a medical practitioner with software projects, particularly open source healthcare software. He co-founded and runs the Leigh Hackspace and offers advice and consultancy in the space between real medical need, agile, and digital. He co-founded Open Health Hub, a non-profit forum about open governance, open source, and open standards in Health IT, and is a member of the Health Informatics Group of the Royal College of General Practitioners.

Datasets

Marcus used a number of datasets both in his work as a GP and for other projects:

- Obesity prevalence
- Growth percentiles from the US CDC
- UK90 LMS Tables

Data Uses

Marcus worked with Obesity prevalence data at a Young Rewired State youth hackday event, creating an API to show how centralised and effective management of such data could help GPs. This involved using the UK90 LMS Tables, which are not open data. (The UK90 LMS tables are used to calculate centiles for weight, height and Body Mass Index, for people aged 4-18yrs in the UK. Between 0 and 4 years the WHO tables are available instead, and they are licensed as open data.) Marcus also wrote a simple website that would calculate centiles.

He also uses data in his work as a GP, in a drive to be objective about patients. This is something Marcus stressed during our conversation:

In order to discuss a child's obesity with their parents, I would prefer to have an objective measure of the degree of obesity; I don't want to risk making a subjective judgement based on the child's appearance. Currently, the gold standard was until recently a (paper) card with a chart, which GPs use to calculate the centile. Most NHS Trusts, and all GPs, have now abandoned the cards (due to cost) and use the freely downloadable PDF versions. Unfortunately these do not integrate at all with digital clinical record systems.

Issues

Marcus complains a lot about formats:

For example, the obesity data is there, but it's in Excel format, and full of medical jargon without relevance to the patient, which defeats the idea of open data; it's a dump of data that is not easily navigable, for example wards are in the dataset multiple times for different centiles and this can be confusing without a proper description of the data

Most datasets require knowledge of medicine, statistical or geographic jargon, and technological competence. Marcus believes these skill sets are only rarely seen in combination, so there should be more effective guidance together with each dataset. This is especially important whenever data is recorded multiple times in the same sheet (for example for multiple wards or age groups), meaning someone who had not noticed this would potentially double or triple-count.

Issues were reported also around access to data that could be open. For example the UK90 LMS tables are owned by the MRC, which regulates access to it; it is in effect a closed dataset, copyrighted and licensed under a non-commercial use agreement. Marcus reports it took him 9 months from applying to receiving a licence agreement and the data:

The MRC didn't charge me for the dataset, but the length of time it took was simply unacceptable.

While discussing with Marcus, an interesting point emerged: the dataset contains public data for public good, so it should be released because the public benefits from it. This is what happens in the US through the CDC, and data should be accompanied by information on how to calculate the centiles.

We are in a so-called obesity epidemic, so why is this dataset not open?. No GP System can calculate centiles. It's like fighting a war without having the right weapons.

Marcus suggests that using such data is vital for a GP to remain objective and avoid forming prejudice about patients.

Whenever QOF data is not available, Marcus suggests the importance for GPs to have some simple unlinked open datasets about prevalence of conditions from other sources. He reports this as fairly restricted at present:

There are some databases you can get access to but they aren't open.

Recommendations

Marcus calls for a wider availability of prevalence tables for other conditions. He suggests this could be achieved by collecting QOF data in a way that does not impinge on GP work and that keeps the public confidence high around privacy issues. In particular, Marcus suggests that consent management systems need to be researched.

Paul Malyon, Data Strategy Manager at Experian

"We use health data in demographics models that benefit our customers' businesses"

Paul Malyon is a Data Strategy Manager at Experian, working on B2B data quality projects. He has worked for Tesco and was a member of the Open Data User Group at Cabinet Office.

Datasets

- ONS NSPD Postcode Directory
- NHS Postcode Directory
- Indices of Multiple Deprivation
- NHS Choices GP Surgeries

Data Uses

Experian sell data validation and quality tools, and consultancy on data products.

For example, they use the NHS Postcode Directory, aiming to answer questions from Clients in the NHS such as: "Do CCGs have the right information about their practices catchment areas?" They used the ONS NSPD to supply NHS bodies with the relevant CCG code for any given postcode, matched to PAF and AddressBase.

Experian have also used a variety of high level health outcome data over the years to answer specific questions from clients in the Health Service. However, due to the nature of many of these datasets, they are not regularly added to demographic products such as Mosaic (a demographic dataset available in a number of countries) due to the impact of decisions on individuals that could be made using this high level modelled data. "For example", reports Paul, "if we were to use the NHS outcome data to assess the likelihood of a smoker living at a property, it would be difficult to make this accurate enough for local authority policy or public health service purposes without additional micro-level data". Using a range of data sources, it helps organisations understand the probable characteristics of an address or postcode area. By matching and modelling data, Mosaic can give the propensity for a range of factors – from household income to their propensity to read a certain national newspaper.

Experian have been doing some work with the ODI to review some of the most commonly used health open datasets. They used the NHS Choices GP Surgery database. Some of the outputs were released as open data by Paul.

Issues

Working on the postcode directories, Paul describes the ONS as open, while the NHS data is not as readily available for onward use. There are also several discrepancies and mismatches. This is data used to state whether a postcode falls within a specific CCG, so the consequences can be serious. There were Scottish CCGs with wrong areas associated – leading to potential confusion for patients and financial errors for the CCGs.

Paul reports: "This issue is now fixed, but there was no defined process to do that officially. I had to find a way, using the right contacts". Experian use the NSPD to help NHS bodies apportion spending on a patient to the correct CCG based upon the home address of that patient. If these addresses are wrong, there are significant efficiency concerns for the health service.

There are too many places to find information, and multiple lists make it difficult to understand which one is canonical. Data flows that generate these datasets are unclear. "For example", Paul asks, "which one is the primary dataset: the one produced by the ONS or the NHS version?" Paul calls for more documentation on how these flows are shaped, in a way that helps identify the master dataset or datasets, and a deeper understanding of what processes compile and update them along with some understanding of how the datasets are used downstream.

Working on the NHS Choices database project Paul reports he found several issues with the files.

We found many, many errors in the file. It's unclear who does the cleaning up and whether anyone in the NHS is monitoring these datasets. Or maybe the generation is crowd-sourced or delegated to the single surgeries. There isn't, either, any mention of standards being used to populate the datasets, which makes it difficult to assess their quality and whether they can be trusted by value-add users like Experian or individual patients.

Recommendations

In terms of business model, Paul suggests that moving to a fully share-alike licence would be problematic for companies like Experian that have a business model based on providing access to curated versions of the data, while the OGL is ok in that respect. However, while the OGL is clearly good for anything relating to administrative information like locations, there is an understanding that different licences might be needed for different categories of data, especially for the ones with important ethical implications. OGL licensed data is used by many different teams in Experian and the Licence is now seen as a mark of 'quality' for open data (even if the quality of the actual data is questionable). This means that new OGL datasets can be used by Experian more quickly than other openly licensed data with the associated benefits passed on to businesses, individuals and the health service in a similarly rapid fashion. However, a word of warning must be issued where the OGL is supplemented with other commercial terms from the likes of the Ordnance Survey – this can create significant delay and cost; sometimes making the data commercially unviable.

Paul has services provision on top of his data wishlist: "From a patient's perspective, the big question would be 'what services are available where and on what date?' This makes an obvious route to follow to produce this dataset, adopt a standard to populate it, and name a custodian."

Tom Smith, Former Managing Director at Oxford Consultants for Social Inclusion (OCSI)

“We provide data insight consultancy that impacts citizens’ lives”

Tom Smith is the Chief Executive of OCSI (until December 2016), and from January 2017 will be MD of the ONS Data Science Campus. He is Chair of the Environment Agency Advisory Group and was a member of the Open Data User Group at Cabinet Office. Tom led the work on the UK government’s Index of Multiple Deprivation (IMD) 2015, and has been involved with developing similar indices using government administrative data since the early 1990s.

OCSI work with a public and community sector clients including Central Government agencies, local authorities, charities, and housing associations. OCSI help these entities to better target their spending and evaluate impact, using models developed from data. As well as producing open data such as the IMD, OCSI use open data in their tools such as Local Insight (local.communityinsight.org).

Datasets

OCSI use both open data and data obtained via Data Sharing Agreements.

Data held by government agencies can give huge insight into social, economic, health and environmental patterns and trends at local level. However, many of these sources cannot be published as open data as they are based on sensitive personal data. In work such as developing the IMD, it is important that data can be securely shared in order to produce robust measures of deprivation for areas across the country.

As examples, some of the datasets used by OCSI are:

- Prescription data at GP Level (open)
- ONS Data on suicide (not open at the level required, but the process to access data is clear)
- Department of Work and Pensions data (not open, accessible, but behind a more convoluted process than ONS)
- Hospital Episode Statistics (not open, but there is an established procedure)
- ONS Population Projections (open)
- Referral, Assessment and Packages of Care (RAP) data for local social care from NHS Digital at Local Authority level (open)

Data Uses

OCSI provides services to several clients based on data insight. They are well known for leading the calculation of the IMD - Indices of Multiple Deprivation, which is made of 37 indicators over 6 areas and a recognised driver for better open data: it allows the development of open data generated from individual-record data. Their Community Insight and Local Insight open data tools, developed in partnership with housing charity HACT, are used by more than 100 local authorities and social housing associations. Another example is their “Planning for Care” project, using ONS population projections together with data on social care needs and services. Working with 30 local authorities, the project was a model of need of social care that allowed local authorities to evaluate actual spending. This answered questions such as “What happens if we change eligibility criteria”, thus enabling LAs to better target their spending and evaluate impacts.

Issues

Tom did not report specific issues.

Recommendations

Tom had two major recommendations.

The first is that data sharing access and agreements should be based on the public benefit of the work being carried out, irrespective of the type of organisation doing the work. The ONS approved researcher scheme is a good example of this, allowing secure access to data for groups who are producing clear public benefit from their analysis.

The second is about the data access process: having a well understood and documented process to securely access data that is not open is hugely important. For example, the NHS Digital agreements for accessing data such as Hospital Episodes Statistics, and the MoJ Justice DataLab, both allow researchers and service providers / commissioners to safely use sensitive data in their research and decisions – while keeping the underlying sensitive data secure.

Tony Hirst, Senior Lecturer at the Open University

"I use open data for educational purposes and to investigate skills gaps in the press industry and in my local area"

Tony Hirst is a Senior Lecturer at the Open University and a well known authority in the industry and the community. He lives on the Isle of Wight. The courses Tony teaches are popular, and they have been ported to MOOC platforms.

Data Uses

The OU have moved from a database course to a data management and analysis course, so that they study the full lifecycle of data, including the legal implications, rather than the theory of relational databases.

My courses aim to answer questions such as How do I look at this data to check if it's insightful? How do I package/share data? How do I create a business out of tractable data problems

Tony also tries to address several research questions. One of the most interesting is whether data can be turned into text for automatic press releases, or how do we use it to detect new signals about outliers. This is aimed mostly at news and media organizations, but he also sees a lot of potential for auto responder chatbots for Care Quality Commission reports and GP complaints.

A data geek in one area could come up with this sort of recipe for their area which also scales as a parameterised function to other areas. If we have national datasets with local relevance, commissioners or journalists can produce reports for local use that can easily parameterised and scaled to other local areas nationally.

Tony is not particularly worried by data correctness or accuracy, as he is more interested in the process of analysing and getting insight of whatever quality.

Data is there to let people ask questions, so quality is not necessarily the main consideration when looking at developing automation processes around the data.

For example, he sees health spend data as a way to understand flows of patients and services. Data about procurement and receipts could be used to analyse symmetries and asymmetries, especially considering that spending between public bodies implicitly gives access to receipts data from other public bodies, not just spending information. He says: "I want to be able to find out whether there are correlations in geographic areas, for example, do all health authorities in my area buy services from another area, and is that area the same?" This could be useful for watchdog organisations.

Issues

The URL on which datasets are published changes frequently, and this creates a lot of broken links. Tony says: "Linking to live datasets in Open University courses is a problem because of likely 'Page not found' errors."

Recommendations

Tony is interested in building a conversation with data publishers: "Give me some questions when you give me data, and let me answer those". He argues that this approach could help build better engagement with users. He sees talking about the internal needs of the NHS as a good starting point for data releases, rather than using data releases to do something outside the system:

We don't know what the NHS needs in terms of data analysis. If we knew, we could develop applications that they find useful.

Tony would welcome engagement with the NHS while the operations are being run, with the goal of letting data analysts offer advice on how to change operations in a more evidence-based way. He has done some research, for example, on appointment attendance, something that could benefit patients and practitioners alike.

Discussion

This report explores the experiences and needs of users of data produced by the health system in England. Our aim is to provide evidence to support the development of policies and programmes concerned with data in the health system rather than to be present a complete view of the data ecosystem.

In this section we present a summary of the findings identifying some key areas, and provide recommendations for each of these key areas. Recommendations should be understood as goals towards which NHS England should work, in collaboration with other organizations where appropriate, in order to encourage interesting and innovative data uses and support the health open data ecosystem.

Who uses data about health?

As outlined above, we did not strictly limit the scope of our research to open data as we found that users frequently combined open, shared and closed data, and it made more sense to research how users actually interact with data rather than present a partial picture.

We found that users of data about health are pragmatic, understanding that much health data can never be open. They feel that having clear processes to access sensitive data and reasonable restrictions over reuse is acceptable if not desirable.

We have identified four broad types of data user:

- Users of data for research purposes, to experiment for academic or professional development reasons
- Users who build business offers or products based on available data
- Users who create evidence to support their line of business
- Users who are medical professionals and use data to audit their procedures.

These uses represent a variety of sectors and businesses, and are similar to the findings of previous work such as the NHS England Open Health Data Project Showcase. However, we found scant evidence of uses of open data by clinical staff working in the NHS. The medical professional with strong data literacy who uses data to improve their clinical practice remains the exception rather than the rule.

How could we help users of data about health?

By far the most pressing issue we discovered was the fragmentation of data publishing locations and variation in standards across the health system. The baffling array of data sources, combined with the complex structure of the NHS creates confusion and extra work for almost anyone doing research, analysis, or building products and services using NHS data.

This problem is exacerbated by the publication of multiple similar datasets, all seemingly canonical, a situation which forces any potential consumer of data to evaluate different sources, often without any good source of information about which should be considered canonical. For example, it is highly confusing, particularly to those without a deep knowledge of the structure of the health system to see multiple subtly different lists of GP surgeries.

A second result of this fragmentation is that many of our interviewees reported that they don't know where to look for data and often use Google as their main entry point. Given that many of the interviewees should be considered expert users of data, often with a professional knowledge of the structure of the NHS, we are forced to consider this a structural failure of the publishing ecosystem within the health system. A common complaint was, once again, that it is hard to identify the primary source of data. Some users still search for information on the old NHS Digital website.

Many users reported a deep frustration with the degree to which the locations at which data is published are not reliable. Datasets are deleted or moved without warning at an alarming frequency, which causes established process both manual and technical to break. In turn, this adds to the work of anyone attempting to use data as a routine activity. Specific examples highlighted were the frequent 'reorganizations' of the NHS England Stats work area which cause old links and navigation structures to change seemingly on a whim, and the major breakage of old links to NHS Digital data publications caused by redesigns and rebranding from NHS Digital. The attitude of NHS Digital to preserving old links was described as " an act of vandalism perpetrated by the state ", indicating the strength of feeling on this issue.

An observation from many users is that there are chunks of data that is not open but widely used, albeit with a restrictive licence. Many of these users would particularly welcome a single, unified, reliable place of publication for datasets from across the health system. This would be particularly effective as a single point of access for both open and shared data - allowing free downloads of open data and encouraging the establishment of open processes to access data that is not open, instead of forcing users to contact different organisations.

Recommendations

Following our interviews, discussions and subsequent analysis, we present two specific recommendations, as well as some general principles for data publishing with concrete examples.

Recommendation 1.

Work towards a single authoritative online point of access for datasets in the health sector, and proceed with a programme of reducing duplication in data publishing. This online service should include both direct access to open data, and a clear process to access shared data.

Recommendation 2.

Establish an expert and reachable data support team with a remit covering the entire health system, not limited to individual institutions. This team should be encouraged to build strong links with the wider data ecosystem and lead a health data user group.

Below we discuss these recommendations in more detail.

A single point of access

All health datasets should be discoverable from a single authoritative online source. There are currently many places where data can be found and downloaded, often with data duplicated several times across the system. We repeatedly discovered this causing significant confusion and wasted effort for data users. Appendix 1 lists more than twenty different services, using different standards, data formats, with wildly different levels of usability.

Many of these are not under direct control of NHS England. Addressing this situation will require collaboration across the system. Establishing a single authoritative source should be combined with a programme of responsibly reducing the duplication of publication currently seen. Teams publishing data in other places should change the location at which they publish datasets, and clearly mark old publication locations as no longer updated, with clear links to the new publication location.

To users the openness of a dataset is not the only relevant feature; they are typically interested in better understanding or explaining a particular domain within health. Shared non-open datasets should be published using the same standards, metadata and discoverability mechanisms as open datasets. Where open datasets have a non-open equivalent (Hospital Episode Statistics for example), publishers should provide clear links explaining how the user would be able to access the non-open version.

Many large data portal services that aggregate data from multiple organizations encounter issues around governance, sustainability and quality. A single portal providing access to all health data would be an ambitious goal towards which to move, and would require careful consideration of these issues.

The processes for publishing data - who can publish, what type of data is contained within this portal, what minimum quality standards are required - should be both clear and strictly enforced. As far as possible checks regarding data quality should be automated: tools such as ODI Data Certificates, JSON Schemas, CSVLint can be built into publishing processes to reduce the burden and introduce some level of objective quality measures.

Publishers who routinely publish data of low quality should be incentivised to improve their processes, and consideration should be given to the establishment of a governance structure able to respond to and improve poor quality data.

The vast majority of open data portals, particularly those that are backed by large organizations are built on open source software. This enables those services to iterate upon their design and functionality, responding to changing user needs over time, with flexibility over vendors and suppliers. This approach aligns with guidance from the Cabinet Office over open standards and open source. Given the maturity and popularity of open source platforms designed to meet these specific needs, this would seem an obvious choice.

Over time, the needs and expectations of both users and publishers will change. Once such a service is live, the business as usual operation should include ongoing user research and usability testing to continuously seek feedback from users to improve the service.

Engaging data users

The experience of health open data users is that the degree and quality of support on offer varies wildly, often as a reflection of organisational and administrative boundaries and priorities. While users frequently spoke highly of the quality of support offered by some institutions, ensuring a consistently high level of service emerged as a clear priority.

Accordingly, there seems to be a need to establish an expert and reachable data support team with a remit covering the entire health system, not limited by institutional or administrative boundaries.

Establishing such a team would present an excellent opportunity to engage with the wider data ecosystem in a proactive way rather than simply reactively responding to support queries.

Although not unanimously seen as positive experiences, the Open Data User Group at Cabinet Office and the Environment Agency Data Advisory Group have helped in different ways their respective organisations understand data issues and prioritise data releases. We advocate a similar form of high-level engagement. It is important to ensure that such groups are weighted strongly towards people with practical recent experience of using data who can provide expertise. The focus of such a group should be on the needs of those who use data, not the needs of policymakers or the institutions that publish data These user groups might be built around themes and be dynamic. For example: a data user group around obesity, dementia, or cancer.

Such a team could also build on recent work by NHS England to showcase the value and impact of health open data. Programmes such as the Open Health Data Project Showcase and the Obesity Data Challenge were generally well received and users reported that such efforts help to reinforce the value of their projects by increasing awareness.

Data Publishing Principles

The open data community has amassed a great deal of knowledge about how to publish data to the web. The degree to which data publishers in the health system apply these best practices has significant variation, and we would particularly like to direct publishers towards the W3C Data on the Web Best Practices, the ODI Open Data Certificates and the ONS Data Publishing Principles.

To complement these, we present some principles and commentary around data publishing. The intended audience for these principles is data publishers within the health system. While we would consider them to be relatively uncontroversial within data publishing circles, they have been written to speak directly to the challenges and issues reported by users of health data. Adherence to these principles by data publishers across the health system would reduce much of the friction reported by data users.

Support data publishers

Although the scope of this report has lead us to focus on the needs of data users, we fully acknowledge that publishers of data will often need support to meet the high standards that are required. Although specific research into their needs would be a separate project we would anticipate that this support should include the following:

- Adopting, producing, and encouraging the use of data standards for data generation and publication, following the lead of widely successful standards such as the Open Contracting Standard.
- Creating effective tools for data publications that allow the publisher to track, monitor, version control, and manage feedback on their datasets.
- Producing clear and publicly available documentation describing the data publication process, fostering engagement and discussion between data producers and consumers.

Data should be discoverable

Users reported significant usability issues with many of the locations in which health data is published.

Many users begin their discovery process using search engines such as Google, and report that much of the data that does exist is difficult to find via searches. Publishing information in such a way that it can be found by Google is a relatively well understood domain, however, and significant advances could be made improving data discovery.

Publishers should refrain from placing open data behind login mechanisms as this prevents discovery.

Publishers should ensure that web pages on which data is published include appropriate metadata and that these pages are easily indexed by third parties . This process can be greatly assisted by taking advantage of data management platforms which allow a QA process that ensures basic metadata is easy to enter, machine readable, and then published in a way that such metadata is exposed to search engines.

Services publishing data should ensure that there are advanced search functionality that enables users to filter the thousands of available datasets consuming appropriate metadata, for example by licence, publisher, and time of publication.

Publishers should optimise for re-use

Users frequently reported technical issues which cause large amounts of duplicated and wasted effort when attempting to re-use data. It is not uncommon for projects to spend 80% of their time cleaning, combining and manipulating data and only 20% of time conducting analysis or visualisation. This chills the effective re-use of health data and reduces the value to the entire health system of publication.

Although we present here a summary of the most pressing needs that emerged in our research, publishers should conduct regular user research with users of data to understand the ongoing needs of these users.

Data Formats

The CSV file format is the most widely adopted by users and it is often the preferred choice even by advanced users. The overwhelming consensus was that CSV is preferable to Excel spreadsheets because it is more portable and comes with a certain expectation that the file is well structured; it is also preferable to sharing linked data, as many users report that “they prefer to do the linking themselves”. All data releases should include raw data as valid CSV. For instance, following the CSV on the Web Charter from the W3C.

For large datasets where CSV files cannot be opened by common spreadsheet software publishers should consider additional publication methods, such as providing APIs for programmatic access to the data.

Data Quality

Much health data is released on a monthly basis, ensuring that timely information is available to the public. However monthly releases can cause significant burden on data users if the publication process does not optimise for re-use. For instance, waiting time statistics for a single month are essentially useless without knowing the figures for previous months, there are vanishingly few real world uses for the snapshot of a single month.

Accordingly publishers should publish periodic and historical data as time series at predictable and stable URLs. The Land Registry Price Paid dataset, published as a single file, as monthly updates, and as annual files would serve as an excellent best-practice example.

Being able to understand what data actually means is a prerequisite to any use of data. For instance, this should include having descriptions of what the fields in a dataset mean, what the possible values, where those values come from. This can be achieved by publishing both documentation and metadata together with each release. Publishers should consider adopting the JSON Table Schema for CSV data to make such information available in a machine readable way that works with existing data tools.

Several users reported licensing problems: in some cases, the licence attached to a given dataset is not clear, the Fingertips website, for example, says that all content is under OGL, but most users report that it is confusing not to see a licence attached to each slice of downloadable data. Publishers should ensure licence information is clearly stated next to the data it pertains to. Again, this is a task where specialist data management software can significantly ease the process.

Published data should be trustworthy

Users repeatedly expressed frustration and confusion around the trustworthiness of published health data. Without the ability to trust a particular data publication, the user is presented with the hard task of assessing the reliability and trustworthiness of a particular dataset before even beginning to use it.

The Office for National Statistics includes a named contact for each publication that users can get in touch with in order to send queries. Similarly, Publishers should ensure that datasets have clear named contacts, or at least a team contact on each dataset, and that the process to provide feedback or obtain support for the dataset is both clear, and provided in the same location as the data.

Publishers should ensure that links to data do not break. It is important to create processes that ensure these datasets will be accessible in the future at the same URL. In conjunction with this, obsolete or superseded data publications should be clearly marked and link to the most recent version provided.

In addition to highlighting the lack of openly available canonical sources of key health datasets, many users reported that they would like key NHS datasets to be managed according to the principles outlined by the Cabinet Office Register Design Authority. We anticipate that this would greatly increase the level of trust in health data infrastructure, and serve as an enabling step that would reduce the cost of building products and services that support the NHS and in conducting research and analysis.

Discussing health registers is beyond the scope of this report, but it would seem a highly worthwhile exercise to establish a working group that explicitly seeks to identify and publish key health data as Registers working with the Register Design Authority.

Candidates identified by our research included:

- The GMC Register
- The NMC Register
- The BCAP Register
- Most data currently administered by the ODS at NHS Digital
- BNF codes

Closing remarks

This report has presented a series of interviews with several professionals working with health datasets in a variety of sectors and roles. The interviewees include users who use data for research purposes, users who build business offers or products based on available data, users who create evidence to support their line of business and users who are medical professionals and use data to audit their procedures.

We offered in this report an analysis of the outcomes of these interviews, identifying common issues in the current data offering and suggestions on what to improve, with a specific attention to the NHS England Data Catalogue.

We developed a set of actionable recommendations to bring these suggestions forward, and believe that the major focus needs to be on raising data capability within the health sector so that data can be used and shared effectively. This will produce better operations, the capability to audit the way services are run, and an overall improvement to patients.

Acknowledgements

We would like to thank the following people for kindly agreeing to being interviewed:

- Anna Powell-Smith
- Callum Tanner
- Edafe Onerhime
- Gavin Jamie
- Jamie Whyte
- Judy Aldred
- Mark Barrett
- Marcus Baw
- Paul Malyon
- Tom Smith
- Tony Hirst

We are also grateful to Paul Downey who provided interesting thoughts and investigation directions on canonical data, to Heather Savory, Mark Braggins and Nick Halliday who kindly offered advice and useful introductions, and to Carl Reynolds and Peter Wells for their comments and feedback on drafts of this report.

The most current version of this document can be found at the URL
http://openhealthcare.org.uk/open-data-in-the-health-sector/#discussion

Appendix 1: Health data services

The table below provides a list of services from which users currently obtain health data.

Name	Institution Responsible	URL
Data.Gov.UK	Cabinet Office	https://data.gov.uk/data/search?theme-primary=Health
NHS Digital	NHS Digital	http://content.digital.nhs.uk/home
NHS Digital Indicator Portal	NHS Digital	https://indicators.hscic.gov.uk/webview/
TRUD	NHS Digital	https://isd.hscic.gov.uk/trud3/user/guest/group/0/home
Data Access Request Service	NHS Digital	http://content.digital.nhs.uk/dars
Organisational Data Service	NHS Digital	https://digital.nhs.uk/organisation-data-service
NHS England Data Catalogue	NHS England	https://data.england.nhs.uk/
NHS England Statistical Work Areas	NHS England	https://www.england.nhs.uk/statistics/statistical-work-areas/
NHS England Resources	NHS England	https://www.england.nhs.uk/resources/ccg-maps/
GP Patient Survey	NHS England	https://gp-patient.co.uk/surveys-and-reports
PHE on gov.uk	PHE	https://www.gov.uk/government/publications/phe-data-and-analysis-tools-a-to-z/phe-data-and-analysis-tools-a-to-z
Public Health Outcomes Framework	PHE	http://www.phoutcomes.info/public-health-outcomes-framework#page/9/gid/1000049/pat/6/par/E12000004/ati/102/are/E06000015
PHE Fingertips	PHE	https://fingertips.phe.org.uk/
National Child and Maternal Health Intelligence Network	PHE	http://www.chimat.org.uk/directory
Local Health Profiles	PHE	http://localhealth.org.uk/

MyNHS	DH	https://www.nhs.uk/service-search/Performance/DownloadData
Northern Ireland GP Datasets website	HSCNI	http://gpdatasets.hscni.net/
HSC Business Services (NI)	HSCNI	http://www.hscbusiness.hscni.net/services/1802.htm
ISD Scotland	ISD	http://www.isdscotland.org/A-to-Z-Index/index.asp
Wales Primary Care Services		http://www.primarycareservices.wales.nhs.uk/data-publications
Dispensing Pharmacy data	NHS BSA	http://www.nhsbsa.nhs.uk/PrescriptionServices/5045.aspx
NHS Staff Survey		http://www.nhsstaffsurveys.com/Page/1021/Past-Results/Historical-Staff-Survey-Results/
Open Exeter		https://digital.nhs.uk/NHAIS/open-exeter
CQC Directory	CQC	http://www.cqc.org.uk/content/how-get-and-re-use-cqc-information-and-data#directory