A-level

SCIENCE IN SOCIETY

SCIS4/PM Case Study of a Scientific Issue

June 2016

Preliminary Material

Instructions

- This Preliminary Material should be opened and issued to candidates on or after 1 May 2016.
- A clean copy of the Preliminary Material will be provided at the start of the Unit 4 examination.

Information

- This Preliminary Material consists of extracts from five sources (**A–E**) on the subject of personal health data.
- This material is being given to you in advance of the Unit 4 examination to enable you to study the content of each extract in preparation for questions based on the material in the examination. Consider the scientific explanations and the ideas about how science works that are involved, as well as the issues raised in the sources.
- You may write notes on this copy of the Preliminary Material, but you will not be allowed to bring this copy, or any other notes you may have made, into the examination room. You will be provided with a clean copy of this Preliminary Material, together with one additional source, **Source F**, at the start of the Unit 4 examination.
- You are not required to carry out any further study of the topic than is necessary for you to gain an understanding of the ideas described and to consider the issues raised. You are not required to understand any detailed **science explanations** beyond those outlined in **Sources A–E** and those in the *Science in Society* specification.
- It is suggested that a minimum of three hours detailed study is spent on this Preliminary Material.

Source A Article taken from ‘Metro’ newspaper, 9 February 2013

Health apps may be spying on us

If you use an app to track your exercise regime or calculate your calorie intake you may have a new symptom to fear.

Personal information that users share with apps such as Map My Fitness, Web MD and iPeriod are being sold to insurance and drug companies, a report has revealed.

Up to 70 third parties harvest data used by people who are tracking diets and even menstrual cycles, said privacy group Evidon.

Jeff Chester, executive director of the Centre for Digital Democracy, said that means some of the ‘most sensitive details of your life’ were available to others.

App companies have denied that the information is personally identifiable and say it is used for site analysis and advertising within apps.

Regulations prevent sale of an individual’s medical records but some US firms have been accused of trying to bypass that by building health profiles from user information on apps.

Andy Kahl, director of data analysis at Evidon, warned some British developers may be doing the same.

‘There’s a mantra I repeat often: if the product is free, then you are the product,’ he added.

END OF SOURCE A

Source B Article taken from The Guardian newspaper, 3 February 2015

‘Public should be consulted on NHS medical data-sharing scheme’

Review of care.data by Nuffield experts says government must ensure people understand their privacy is at risk

Ian Sample

Plans to combine NHS patient records into a national database must be reconsidered by the government to ensure that people understand their privacy cannot be guaranteed, leading experts say.

The government’s care.data scheme involves the creation of a database that holds anonymised patient records and information on hospital admissions. It was due to start last year, but was postponed for six months amid concerns over consent and privacy.

The plan was launched as an opt-out scheme that assumed people were happy to have their records used for medical research and by life science companies unless they specifically opted out.

However, the NHS has disregarded tens of thousands of requests by patients to opt out of the health service’s system of sharing medical records. Officials admitted last month that not sharing the data would affect the treatment patients received, such as cancer screening services.

On Tuesday, an expert panel convened by the Nuffield Council on Bioethics said the public should be consulted on the best way to run the plan and whether consent through opting in was more appropriate.

It added that information leaflets describing care.data should stress that there was no way of ensuring that sensitive information from patients’ records would never be released into the public domain, either by hackers or from people abusing the system.

When care.data was launched, millions of homes received leaflets explaining the benefits of the plan. It recommended people contact their GP if they wished to opt out. But Prof Martin Richards, chair of the Nuffield working party on the ethics of big data said the leaflet “should state that anonymising patient records cannot guarantee privacy”.

In a report on the collection, linking and use of data in biomedical research and healthcare, published on Tuesday, the panel argues for prison sentences to punish people who intentionally leak information from patients’ records, whether the leaks cause harm or distress or not.

The panel conducted the review in light of a flurry of major projects that bring together medical and health data from huge numbers of patients, including care.data, the 100,000 genomes project, UK Biobank and the Scottish Informatics Programme.

“We now generate more health and biological data than ever before,” said Prof Richards. “This includes GP records, laboratory tests, clinical trials and health apps, and it has become cheaper to collect, store and analyse this data.”

He said the public had a strong interest in the benefits of using medical data to further knowledge and improve healthcare, but that they still had privacy concerns. “If we don’t get this right, we risk losing public trust in research, and ultimately missing out on the benefits this type of research can bring,” he added.

Turn over ▶

The report urged health authorities to track the use of patient data so they could provide people with complete audit trails of who used their medical data and for what reasons. People should also have a say in how their data will be used, the committee said.

Susan Wallace, a member of the panel, added: “Any data project should first take steps to find out how people expect their data to be used and engage with those expectations through a process of continued participation and review.”

END OF SOURCE B

Source C Press release from Bristol University, 2 April 2014

1,000th paper for world-leading health study

One of the world's largest population studies, which collects vast amounts of data from 32,000 participants to give new insights into our health, is today celebrating an important landmark in its acclaimed research history.

Researchers from Children of the 90s have published their 1,000th paper today, showing that men who started smoking before the age of 11 had fatter sons.

It's the latest finding from the project, officially called the Avon Longitudinal Study of Parents and Children (ALSPAC), which is based at the University of Bristol and has provided data to almost 600 academics worldwide.

In 1991 and 1992, more than 14,000 pregnant women in Bristol and the surrounding area agreed to take part in a ground-breaking study that would follow them and their babies, recording information to be used by scientists to investigate ways in which the environment and genetics interact over time to influence health and development.

Unlike many other studies, ALSPAC recruited women during pregnancy so that precise information could be gathered about the children's lives even before they were born. It has since followed their development through to adulthood and beyond.

Today, it has 32,000 participants including the original mums, their 14,500 children, 3,000 dads, 200 'children of the children' and 550 siblings. Even though 500 participants now live abroad, they still keep in touch with the research team.

The project has just received almost £8 million in core funding from the Medical Research Council (MRC) and the Wellcome Trust to continue its work until March 2019.

Teasing out the relationships and interactions between environmental factors and individual genotypes is extremely difficult, and requires vast amounts of data for researchers to study – which is exactly what ALSPAC has collected from its committed families.

To date, 1.4 million biological samples - including urine, hair, blood and DNA - have been collected alongside the answers to 500,000 questions and detailed records of everyday characteristics such as diet, lifestyle, socioeconomic status and emotional health.

So rich is the ALSPAC resource that the World Health Organization and the US National Institutes of Health are formally referring to the study as a model for researchers wishing to set up similar birth cohort studies worldwide.

Whether it's advice to safeguard an unborn child, or the benefits of exercise or the dangers of drugs or alcohol, ALSPAC research has given the world a wealth of practical wisdom that millions of people now put into practice every day.

Some of the key research findings from the project include:

- mothers who consume less fish during pregnancy have children with lower IQs and impaired ability to focus their eyes in early childhood
- depression in fathers is associated with adverse emotional and behavioural outcomes in children aged 3.5 years and an increased risk of conduct problems in boys

Turn over ▶

- children brought up in very hygienic homes are more likely to develop asthma
- children from privileged backgrounds tend to be taller and thinner than those from other families but they also may have weaker bones
- smoking cannabis during pregnancy can result in a smaller baby
- peanut allergies may be linked to the use of certain skin creams.

END OF SOURCE C

Source D Article from Financial Times, 28 March 2014

Big data: are we making a big mistake?

By Tim Harford

Big data is a vague term for a massive phenomenon that has rapidly become an obsession with entrepreneurs, scientists, governments and the media

Five years ago, a team of researchers from Google announced a remarkable achievement in one of the world's top scientific journals, Nature. Without needing the results of a single medical check-up, they were nevertheless able to track the spread of influenza across the US. What's more, they could do it more quickly than the Centers for Disease Control and Prevention (CDC). Google's tracking had only a day's delay, compared with the week or more it took for the CDC to assemble a picture based on reports from doctors' surgeries. Google was faster because it was tracking the outbreak by finding a correlation between what people searched for online and whether they had flu symptoms.

Not only was 'Google Flu Trends' quick, accurate and cheap, it was theory-free. Google's engineers didn't bother to develop a hypothesis about what search terms – 'flu symptoms' or 'pharmacies near me' – might be correlated with the spread of the disease itself. The Google team just took their top 50 million search terms and let the algorithms do the work.

The success of Google Flu Trends became emblematic of the hot new trend in business, technology and science: 'Big Data'. What, excited journalists asked, can science learn from Google?

As with so many buzzwords, 'big data' is a vague term, often thrown around by people with something to sell. Some emphasise the sheer scale of the data sets that now exist – the Large Hadron Collider's computers, for example, store 15 petabytes a year of data, equivalent to about 15,000 years' worth of your favourite music.

But the 'big data' that interests many companies is what we might call 'found data', the digital exhaust of web searches, credit card payments and mobiles pinging the nearest phone mast. Google Flu Trends was built on found data and it's this sort of data that interests me here. Such data sets can be even bigger than the Large Hadron Collider data – Facebook's is – but just as noteworthy is the fact that they are cheap to collect relative to their size, they are a messy collage of datapoints collected for disparate purposes and they can be updated in real time. As our communication, leisure and commerce have moved to the internet and the internet has moved into our phones, our cars and even our glasses, life can be recorded and quantified in a way that would have been hard to imagine just a decade ago.

Cheerleaders for big data have made four exciting claims, each one reflected in the success of Google Flu Trends: that data analysis produces uncannily accurate results; that every single data point can be captured, making old statistical sampling techniques obsolete; that it is passé to fret about what causes what, because statistical correlation tells us what we need to know; and that scientific or statistical models aren't needed because, to quote 'The End of Theory', a provocative essay published in Wired in 2008, "with enough data, the numbers speak for themselves".

Unfortunately, these four articles of faith are at best optimistic oversimplifications. At worst, according to David Spiegelhalter, Winton Professor of the Public Understanding of Risk at Cambridge University, they can be "Absolute nonsense."

Turn over ▶

Found data underpin the new internet economy as companies such as Google, Facebook and Amazon seek new ways to understand our lives through our data exhaust. Since Edward Snowden's leaks about the scale and scope of US electronic surveillance it has become apparent that security services are just as fascinated with what they might learn from our data exhaust, too.

Consultants urge the data-naive to wise up to the potential of big data. A recent report from the McKinsey Global Institute reckoned that the US healthcare system could save $300bn a year – $1,000 per American – through better integration and analysis of the data produced by everything from clinical trials to health insurance transactions to smart running shoes.

But while big data promise much to scientists, entrepreneurs and governments, they are doomed to disappoint us if we ignore some very familiar statistical lessons.

"There are a lot of small data problems that occur in big data," says Spiegelhalter. "They don't disappear because you've got lots of the stuff. They get worse."

Four years after the original Nature paper was published, Nature News had sad tidings to convey: the latest flu outbreak had claimed an unexpected victim: Google Flu Trends. After reliably providing a swift and accurate account of flu outbreaks for several winters, the theory-free, data-rich model had lost its nose for where flu was going. Google's model pointed to a severe outbreak but when the slow-and-steady data from the CDC arrived, they showed that Google's estimates of the spread of flu-like illnesses were overstated by almost a factor of two.

The problem was that Google did not know – could not begin to know – what linked the search terms with the spread of flu. Google's engineers weren't trying to figure out what caused what. They were merely finding statistical patterns in the data. They cared about correlation rather than causation. This is common in big data analysis. Figuring out what causes what is hard (impossible, some say). Figuring out what is correlated with what is much cheaper and easier. That is why, according to Viktor Mayer-Schönberger and Kenneth Cukier's book, *Big Data*, "causality won't be discarded, but it is being knocked off its pedestal as the primary fountain of meaning".

But a theory-free analysis of mere correlations is inevitably fragile. If you have no idea what is behind a correlation, you have no idea what might cause that correlation to break down. One explanation of the Flu Trends failure is that the news was full of scary stories about flu in December 2012 and that these stories provoked internet searches by people who were healthy. Another possible explanation is that Google's own search algorithm moved the goalposts when it began automatically suggesting diagnoses when people entered medical symptoms.

Google Flu Trends will bounce back, recalibrated with fresh data – and rightly so. There are many reasons to be excited about the broader opportunities offered to us by the ease with which we can gather and analyse vast data sets. But unless we learn the lessons of this episode, we will find ourselves repeating it.

Statisticians have spent the past 200 years figuring out what traps lie in wait when we try to understand the world through data. The data are bigger, faster and cheaper these days – but we must not pretend that the traps have all been made safe. They have not.

In 1936, the Republican Alfred Landon stood for election against President Franklin Delano Roosevelt. The respected magazine, The Literary Digest, shouldered the responsibility of forecasting the result. It conducted a postal opinion poll of astonishing ambition, with the aim of reaching 10 million people, a quarter of the electorate. The deluge of mailed-in replies can hardly be imagined but the Digest seemed to be relishing the scale of the task. In late August it reported, “Next week, the first answers from these ten million will begin the incoming tide of marked ballots, to be triple-checked, verified, five-times cross-classified and totalled.”

After tabulating an astonishing 2.4 million returns as they flowed in over two months, The Literary Digest announced its conclusions: Landon would win by a convincing 55 per cent to 41 per cent, with a few voters favouring a third candidate.

The election delivered a very different result: Roosevelt crushed Landon by 61 per cent to 37 per cent. To add to The Literary Digest’s agony, a far smaller survey conducted by the opinion poll pioneer George Gallup came much closer to the final vote, forecasting a comfortable victory for Roosevelt. Mr Gallup understood something that The Literary Digest did not. When it comes to data, size isn’t everything.

Opinion polls are based on samples of the voting population at large. This means that opinion pollsters need to deal with two issues: sample error and sample bias.

Sample error reflects the risk that, purely by chance, a randomly chosen sample of opinions does not reflect the true views of the population. The ‘margin of error’ reported in opinion polls reflects this risk and the larger the sample, the smaller the margin of error. A thousand interviews is a large enough sample for many purposes and Mr Gallup is reported to have conducted 3,000 interviews.

But if 3,000 interviews were good, why weren’t 2.4 million far better? The answer is that sampling error has a far more dangerous friend: sampling bias. Sampling error is when a randomly chosen sample doesn’t reflect the underlying population purely by chance; sampling bias is when the sample isn’t randomly chosen at all. George Gallup took pains to find an unbiased sample because he knew that was far more important than finding a big one.

The Literary Digest, in its quest for a bigger data set, fumbled the question of a biased sample. It mailed out forms to people on a list it had compiled from automobile registrations and telephone directories – a sample that, at least in 1936, was disproportionately prosperous. To compound the problem, Landon supporters turned out to be more likely to mail back their answers. The combination of those two biases was enough to doom The Literary Digest’s poll. For each person George Gallup’s pollsters interviewed, The Literary Digest received 800 responses. All that gave them for their pains was a very precise estimate of the wrong answer.

The big data craze threatens to be The Literary Digest all over again. Because found data sets are so messy, it can be hard to figure out what biases lurk inside them – and because they are so large, some analysts seem to have decided the sampling problem isn’t worth worrying about. It is.

Professor Viktor Mayer-Schönberger of Oxford’s Internet Institute, co-author of *Big Data*, told me that his favoured definition of a big data set is one where ‘N = All’ – where we no longer have to sample, but we have the entire background population. Returning officers do not estimate an election result with a representative tally: they count the votes – all the votes. And when ‘N = All’ there is indeed no issue of sampling bias because the sample includes everyone.

But is ‘N = All’ really a good description of most of the found data sets we are considering? Probably not. “I would challenge the notion that one could ever have all the data,” says Patrick Wolfe, a computer scientist and professor of statistics at University College London.

Turn over ▶

An example is Twitter. It is in principle possible to record and analyse every message on Twitter and use it to draw conclusions about the public mood. (In practice, most researchers use a subset of that vast 'fire hose' of data.) But while we can look at all the tweets, Twitter users are not representative of the population as a whole. (According to the Pew Research Internet Project, in 2013, US-based Twitter users were disproportionately young, urban or suburban, and black.)

There must always be a question about who and what is missing, especially with a messy pile of found data. Kaiser Fung, a data analyst and author of *Numbersense*, warns against simply assuming we have everything that matters. "N = All is often an assumption rather than a fact about the data," he says.

Consider Boston's *Street Bump* smartphone app, which uses a phone's accelerometer to detect potholes without the need for city workers to patrol the streets. As citizens of Boston download the app and drive around, their phones automatically notify City Hall of the need to repair the road surface. Solving the technical challenges involved has produced, rather beautifully, an informative data exhaust that addresses a problem in a way that would have been inconceivable a few years ago. The City of Boston proudly proclaims that the "data provides the City with real-time information it uses to fix problems and plan long term investments."

Yet what *Street Bump* really produces, left to its own devices, is a map of potholes that systematically favours young, affluent areas where more people own smartphones. *Street Bump* offers us 'N = All' in the sense that every bump from every enabled phone can be recorded. That is not the same thing as recording every pothole. As Microsoft researcher Kate Crawford points out, found data contain systematic biases and it takes careful thought to spot and correct for those biases. Big data sets can seem comprehensive but the 'N = All' is often a seductive illusion.

Who cares about causation or sampling bias, though, when there is money to be made? Corporations around the world must be salivating as they contemplate the uncanny success of the US discount department store Target, as famously reported by Charles Duhigg in The New York Times in 2012. Duhigg explained that Target has collected so much data on its customers, and is so skilled at analysing that data, that its insight into consumers can seem like magic.

Duhigg's killer anecdote was of the man who stormed into a Target near Minneapolis and complained to the manager that the company was sending coupons for baby clothes and maternity wear to his teenage daughter. The manager apologised profusely and later called to apologise again – only to be told that the teenager was indeed pregnant. Her father hadn't realised. Target, after analysing her purchases of unscented wipes and magnesium supplements, had.

Statistical sorcery? There is a more mundane explanation.

"There's a huge false positive issue," says Kaiser Fung, who has spent years developing similar approaches for retailers and advertisers. What Fung means is that we didn't get to hear the countless stories about all the women who received coupons for babywear but who weren't pregnant.

Hearing the anecdote, it's easy to assume that Target's algorithms are infallible – that everybody receiving coupons for onesies and wet wipes is pregnant. This is vanishingly unlikely. Indeed, it could be that pregnant women receive such offers merely because everybody on Target's mailing list receives such offers. We should not buy the idea that Target employs mind-readers before considering how many misses attend each hit.

In Charles Duhigg's account, Target mixes in random offers, such as coupons for wine glasses, because pregnant customers would feel spooked if they realised how intimately the company's computers understood them.

Fung has another explanation: Target mixes up its offers not because it would be weird to send an all-baby coupon-book to a woman who was pregnant but because the company knows that many of those coupon-books will be sent to women who aren't pregnant after all.

None of this suggests that such data analysis is worthless: it may be highly profitable. Even a modest increase in the accuracy of targeted special offers would be a prize worth winning. But profitability should not be conflated with omniscience.

In 2005, John Ioannidis, an epidemiologist, published a research paper with the self-explanatory title, 'Why Most Published Research Findings Are False'. The paper became famous as a provocative diagnosis of a serious issue. One of the key ideas behind Ioannidis's work is what statisticians call the 'multiple-comparisons problem'.

It is routine, when examining a pattern in data, to ask whether such a pattern might have emerged by chance. If it is unlikely that the observed pattern could have emerged at random, we call that pattern 'statistically significant'.

The multiple-comparisons problem arises when a researcher looks at many possible patterns. Consider a randomised trial in which vitamins are given to some primary schoolchildren and placebos are given to others. Do the vitamins work? That all depends on what we mean by "work". The researchers could look at the children's height, weight, prevalence of tooth decay, classroom behaviour, test scores, even (after waiting) prison record or earnings at the age of 25. Then there are combinations to check: do the vitamins have an effect on the poorer kids, the richer kids, the boys, the girls? Test enough different correlations and fluke results will drown out the real discoveries.

There are various ways to deal with this but the problem is more serious in large data sets, because there are vastly more possible comparisons than there are data points to compare. Without careful analysis, the ratio of genuine patterns to spurious patterns – of signal to noise – quickly tends to zero.

Worse still, one of the antidotes to the multiple-comparisons problem is transparency, allowing other researchers to figure out how many hypotheses were tested and how many contrary results are languishing in desk drawers because they just didn't seem interesting enough to publish. Yet found data sets are rarely transparent. Amazon and Google, Facebook and Twitter, Target and Tesco – these companies aren't about to share their data with you or anyone else.

New, large, cheap data sets and powerful analytical tools will pay dividends – nobody doubts that. And there are a few cases in which analysis of very large data sets has worked miracles. David Spiegelhalter of Cambridge points to Google Translate, which operates by statistically analysing hundreds of millions of documents that have been translated by humans and looking for patterns it can copy. This is an example of what computer scientists call 'machine learning', and it can deliver astonishing results with no preprogrammed grammatical rules. Google Translate is as close to a theory-free, data-driven algorithmic black box as we have – and it is, says Spiegelhalter, "an amazing achievement". That achievement is built on the clever processing of enormous data sets.

But big data do not solve the problem that has obsessed statisticians and scientists for centuries: the problem of insight, of inferring what is going on, and figuring out how we might intervene to change a system for the better.

Turn over ▶

“We have a new resource here,” says Professor David Hand of Imperial College London. “But nobody wants ‘data’. What they want are the answers.”

To use big data to produce such answers will require large strides in statistical methods.

“It’s the wild west right now,” says Patrick Wolfe of UCL. “People who are clever and driven will twist and turn and use every tool to get sense out of these data sets, and that’s cool. But we’re flying a little bit blind at the moment.”

Statisticians are scrambling to develop new methods to seize the opportunity of big data. Such new methods are essential but they will work by building on the old statistical lessons, not by ignoring them.

Recall big data’s four articles of faith. Uncanny accuracy is easy to overrate if we simply ignore false positives, as with Target’s pregnancy predictor. The claim that causation has been “knocked off its pedestal” is fine if we are making predictions in a stable environment but not if the world is changing (as with Flu Trends) or if we ourselves hope to change it. The promise that ‘N = All’, and therefore that sampling bias does not matter, is simply not true in most cases that count. As for the idea that ‘with enough data, the numbers speak for themselves’ – that seems hopelessly naive in data sets where spurious patterns vastly outnumber genuine discoveries.

‘Big data’ has arrived, but big insights have not. The challenge now is to solve new problems and gain new answers – without making the same old statistical mistakes on a grander scale than ever.

END OF SOURCE D

Source E

Preventive Medicine 70 (2015) 17–18

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Commentary

A “big data” approach to HIV epidemiology and prevention

Sean D. Young *

Department of Family Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

ARTICLE INFO

Available online 8 November 2014

Keywords:
Digital epidemiology
Social media
HIV prevention
Big data

ABSTRACT

The recent availability of “big data” from social media and mobile technologies provides promise for development of new tools and methods to address the HIV epidemic. This manuscript presents recent work in this growing area of bioinformatics, digital epidemiology, and disease modeling, describes how it can be applied to address HIV prevention, and presents issues that need to be addressed prior to implementing a mobile technology big-data approach to HIV prevention.

Although HIV remains a tremendous public health challenge after 3 decades of prevention and treatment efforts, the recent availability of “big data” from new technologies provides promise for the development of new tools and methods to address the HIV epidemic.

In 2011, it was estimated that more than 1.1 million people were living with HIV/AIDS and 50,000 people were newly diagnosed with HIV in the United States (CDC, 2013). The challenge to combat the spread of HIV is particularly salient among men who have sex with men (MSM), as in 2010, more than half of newly diagnosed HIV cases were among MSM. Traditional public health strategies struggle to reach MSM, leading MSM to be less likely to be tested for HIV, access and be retained in care, adhere to treatment, and survive 5 years after diagnosis (Bogart et al., 2010; CDC, 2002; Hall et al., 2007). Innovative strategies are needed to provide new tools and better methods of disease surveillance to improve HIV prevention and treatment and reduce the disparities among all populations affected by HIV.

The flood of “big data” from mobile technologies, such as social media, mobile phones, and mobile applications, provides the promise to be able to use these data to develop new HIV monitoring and epidemiology methods, and to provide insights on how to improve HIV interventions and respond to disease outbreaks. Because of the large and increasing use of mobile technologies among African Americans, Latinos, and gay populations (Smith, 2010; Young, 2012), analyses of big data from mobile technologies might be particularly helpful in addressing HIV prevention and treatment efforts among these high-risk populations (Young and Jaganath, 2013).

Although there is no clear definition, “big data” refers to datasets that are often characterized by their enormity and complexity (Grant, 2012). These large datasets are available because affordable and easy-to-use technologies have increased the ability for public health researchers to generate large amounts of data (Grant, 2012; Lohr, 2012; Marx, 2013; Murdoch and Detsky, 2013). For example, the Human Genome Project (HGP), completed in 2003, was an international collaboration to sequence all the base pairs in the human genome. Individual labs were tasked to contribute data from certain areas of the human genome to the HGP database. The combination of these data and the additional combined data have made HGP a classic example of big data (genome.gov, 2014). Big data contain not only relational (structured) data that are conventional in most medical and quantitative datasets, but also unstructured (often free-text) data that can be useful for secondary analyses and qualitative epidemiologic measures (Murdoch and Detsky, 2013). Unstructured data are important in health research because we can use these free-text data to draw inferences about real-time behaviors and sentiments (Lohr, 2012; Young et al., 2014). For example, social media sites and search engines can be used to collect unstructured posts, messages, searches, updates, and tweets from their users and use these data to inform future public health outbreaks. Influenza researchers have used these unstructured social media data (e.g., from Google searches and Tweets) to predict influenza patterns ahead of the Centers for Disease Control and Prevention (CDC) to strengthen public health preparedness (Broniatowski et al., 2013; Ginsberg et al., 2009; Polgreen et al., 2008).

In fact, the majority of work in this area to date has focused on using big data to respond to influenza outbreaks. For example, Google Flu Trends was designed to tally the number of search terms at any given time that were associated with influenza.

* Center for Digital Behavior, Department of Family Medicine, University of California at Los Angeles, 10880 Wilshire Blvd, Suite 1800, USA. Fax: +1 310 794 3580.

http://dx.doi.org/10.1016/j.ypmed.2014.11.002
0091-7435/© 2014 Elsevier Inc. All rights reserved.

Turn over ▶

The Google Flu Trends algorithm looked at searches for terms such as "influenza" and "early signs of the flu" in order to determine whether these search terms could be used to monitor cases of influenza. Using data from the CDC's surveillance system (ILINet), a consolidated database of influenza cases reported by the CDC, state and local health departments, and health care providers, studies have found a high correlation (> .9) between Google Flu Trends and ILINet (Cook et al., 2011), suggesting the potential for big data bioinformatics approaches such as Google Flu Trends in monitoring influenza outbreaks. Because of the fairly open access to conversations on Twitter through the "Twitterhose" (Young et al., 2014) Aramaki et al. (2011) applied a similar approach looking at Tweet data (from Twitter) in Japan that included keywords associated with flu-like illness (e.g., cough, fever, and chills). They found that these tweets had up to a .97 correlation with reported influenza cases in Japan.

Although these approaches might apply to a broad number of public health topics, such as influenza, diabetes prevention and management, substance abuse, and sexual health, there has been limited or no work that has been conducted on these topics. Research has been conducted on this topic around HIV epidemiology and prevention, and analysis of social media big data appears to be feasible for use in that area. After filtering tweets for HIV risk-related keywords and phrases suggesting the occurrence of present or future HIV risk (e.g. sexual behaviors and drug use behaviors), researchers found a high correlation between the geography of these County-level HIV-related tweets and actual CDC reported HIV cases (Young et al., 2014). This study provided further evidence that social media data have the potential to provide a more cost-effective and real-time alternative for HIV remote monitoring and surveillance. Social media data have also aided researchers in HIV prevention efforts, such as HIV interventions and the ability to distribute home-HIV testing kits to those in need. After analyzing free-text posts from an online community focused on HIV prevention, one study found that individuals who posted about HIV prevention and testing, compared to those who posted about other topics, wound up being significantly more likely to request an HIV self-testing kit (Young and Jaganath, 2013). These types of insights based on social media could be valuable in providing health departments with information on how many tests or prevention products might be needed, and determining real-time information on where those health services are about to be requested. More research is needed to refine the methods of using big data in HIV as well as other areas of public health, providing an important and necessary opportunity for HIV and public health researchers.

There has already been criticism about some of the current methods of using technology data for monitoring health outcomes, including the reliability and validity of the data and methods (Lazer et al., 2014), making it important to highlight issues that need to be addressed prior to using big data from technologies for HIV monitoring. First, there are usability issues with big data approaches as many government agencies, local organizations, and even academic public health departments currently lack the infrastructure to handle big data (Grant, 2012; Murdoch and Detsky, 2013). Traditional statistical infrastructure is not powerful enough to address the complexity of big data and unstructured data (Grant, 2012; Murdoch and Detsky, 2013). Instead, collaborations between public health researchers and computer scientists trained in machine learning/data mining are encouraged and perhaps essential to provide the necessary infrastructure for storing and analyzing big data. Second, HIV data need to be released and updated frequently in order to better develop methods of using big data to monitor HIV cases. For example, in the HIV twitter study mentioned above, although tweets were retrieved in real-time during 2012, the most current easily accessible HIV data were from cases in 2009. Therefore, the study could not determine whether social media might be used to monitor ongoing and future HIV cases, but rather could only determine an association between tweets and historical HIV cases (Young et al., 2014). Although there might be limited changes in HIV prevalence from one year to the next, providing access to frequently updated data on HIV cases could help to improve statistical models designed for public health departments to monitor and respond to HIV cases.

This manuscript provides a call for researchers to use technology big data and explore how they can be used to develop new methods of monitoring HIV transmission and other public health concerns. Refinement of these digital epidemiology or bioinformatics methods will help to facilitate the transition from research to practice so that public health organizations can more readily incorporate these approaches into their epidemiology prevention and monitoring efforts.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

Acknowledgments

We wish to thank the National Institute of Mental Health (NIMH) (K01 MH 090884) for funding this work.

This paper contained 19 references which have been removed.

END OF SOURCE E

There are no sources printed on this page

There are no sources printed on this page

Copyright information

For confidentiality purposes, from the November 2015 examination series, acknowledgements of third party copyright material will be published in a separate booklet rather than including them on the examination paper or support materials. This booklet is published after each examination series and is available for free download from www.aqa.org.uk after the live examination series.

Permission to reproduce all copyright material has been applied for. In some cases, efforts to contact copyright-holders may have been unsuccessful and AQA will be happy to rectify any omissions of acknowledgements. If you have any queries please contact the Copyright Team, AQA, Stag Hill House, Guildford, GU2 7XJ.