Lots of Lovely Numbers, but Why Does Everyone Make It So Difficult?

John Pullinger1,∗

1International Association for Official Statistics, the Hague, the Netherlands
∗Correspondence: johnpullingerlimited@gmail.com
https://doi.org/10.1016/j.patter.2020.100033

Based on keynote at the 4th International Conference on Administrative Data, Cardiff, UK, December 10, 2019.

When I was starting out on research projects in the late 1970s, the biggest problem was data. Data were a scarce resource, and we had to make every observation count. I was lucky because my main data source was the census that was, and remains, the most wonderful quarry for insight into what people’s lives are like in countries around the world. In this year, the central year of the United Nations “2020 Census round” it is right to celebrate censuses and the great act of citizenship that participation in the census represents.

However, since then the data landscape has been transformed. The number and richness of data sources, especially administrative data sources, would have been inconceivable to my younger self. At the same time, the demand for good research to help answer urgent questions is growing rapidly. Can you help us understand why our economies are flattening and too many people are getting left behind? How do we focus efforts to fight injustice and hold the powerful to account? What is really happening to our climate, and how can we create a sustainable future for our children and grandchildren? How can we make our communities safe and secure?

Lots of lovely numbers. Lots of demand to use them to create new insights for decision making. But why does everyone make it so difficult?

Respect

Research is not a technical activity. It is political. As has always been true, and will continue to be true, information is power. Power can be wielded for public good or to further vested interest. And statistical research can serve citizens and support democracy or cloak a political claim in a veneer of respectability and hoodwink people into a fantasy that ultimately only serves those who already wield power. It is no wonder people are skeptical when any one of us says, “Give us your data and we will do good, good things with it.” To succeed, we must show respect, earn trust, and differentiate what we do as something to be valued. This is not about being difficult; it is an essential part of every researcher’s license to operate.

As data scientists, we are working with information that people reasonably expect to be kept confidential. Data protection in general and, in Europe, the General Data Protection Regulation are enablers rather than barriers to analysis. Their protocols and codes are fair dealing enshrined in law. In the case of businesses, if our promise to protect their secrets is not honored, they could be out of business.

Linked to data protection is the need to support a strong legal framework. The ability to sanction breaches of the law is especially important in earning respect, particularly for international projects. Getting it right can take time and can even mean that some projects cannot proceed. But getting it wrong can jeopardize trust across the board.

Increasingly our concerns about data enter the complex world of ethics. As technology reduces the limitations on what we can do with data, we need to address the question of what we should be doing. The growing capability to use machine learning algorithms to refine sophisticated linkages across data sources increases the risk of crossing a line.

Technology change also highlights security risks that need to be managed well if the research community is to deserve respect. Security used to be about physical barriers around sensitive material and not losing discs in the post; now it is also about the fast-shifting world of cyber security. It can feel like another hurdle, but understanding encryption and other protections is another supporting part of the research toolkit for trust.

Respect requires an appreciation of what people care about. Citizens have particular anxieties about reuse of data held by the State. Data protection, a strong legal framework, ethical action, and effective security are necessary conditions for the research use of modern data sources, but the most critical and fundamental test is public acceptability.

And the issue of respect is not just about data subjects. The researcher also needs to show respect to data custodians. We cannot just rock up and say, “I’m a researcher doing good stuff, give me your data.” We have to demonstrate to them that our plans are compliant with their legal and other concerns, that we recognize and minimize the time it will take them to work through with us any necessary agreements and protocols, and that we realize that their systems are designed for another purpose and that there are costs, especially IT costs, associated with preparation and extraction of the data we need.

Quality

“OK,” I hear you say, “Getting respect isn’t really about people making things difficult. It is about getting a license to operate. It is part of my job. But once I...
have done all that and got hold of the data, people still give me a load of grief. Why can’t I just get on with my analysis?”

The great thing about administrative data is that they give 100% data compared to the results from only a survey. They can be real time. They are usually highly granular. They are ready made. That’s great surely? I can only muster a qualified “maybe” on that one. And it’s not me being difficult again.

The first question to ask is 100% of what? It may well not be the population group of interest to you. In the UK, an iconic example is the administrative count of the number of people claiming unemployment benefits. A great data source. The trouble was that during the 1980s the government frequently changed the definition of who can make a claim and the number representing 100% became progressively smaller. Despite heroic efforts from statisticians to produce figures on a consistent basis, the numbers derived from this administrative data source became more and more removed from people’s experience of the labor market, and the public lost trust in what was widely perceived as manipulation of the figures.

Scratch the surface of an administrative dataset and you will often find things that need to be considered before using it to research particular questions. An example would be looking at entry and exit data for people traveling into and out of the country to investigate immigration. How do you take account of those people who quite legally have more than one passport and enter using one and leave using another? How do you count people in and count them out again if they use multiple identities? There are inherent quality issues in administrative data that require a careful understanding of what each variable represents.

Another challenge is that a national administrative dataset appears to be consistent for each part of the country covered, but unlike surveys, there may be only limited efforts to ensure that coverage is measured. Datasets I have used in the past such as electoral data or GP registration data can be impacted, where a particularly effective electoral registration officer gets much higher coverage than their peers or poor incentives to remove patients who have moved result in differentially inflated lists. If these administrative data sources are used to assist in the estimation of the population and these population estimates are used to assess allocation of resources between areas, then this kind of geographical disparity has to be well understood and adjustments made.

Administrative data are also subject to questions about consistency over time. A recent example is statistics on divorces. The headlines told of substantial reductions in the divorce rate. The small print mentioned that problems in the divorce courts had meant that cases were being delayed and had not yet shown up in the figures. When we know that analysis is going to be used to tell a story about change over time, this kind of thing needs to be avoided.

For some data there is the double challenge of consistency over space and time. A particularly tricky source is police-recorded crime. Each force administers its own data, and following widespread concerns about underreporting, some forces moved more quickly than others to improve. The researcher has to assess quality against this background if any sensible conclusions are to be drawn. Is this crime type in this area going up because of something happening on the ground or from better reporting? Often this means triangulation against other data sources to assess the true picture. In the case of violent crime, for example, to analyze police-recorded crime data in conjunction with survey data from victims and hospital data on injuries. One source alone is not enough to reveal valuable insights.

Often the issues come down to what is going on at the point of data collection. What are the incentives on police officers, teachers, or nurses to record data in a particular way (or at all)? In situations where there are targets, the impact can be striking. Hospital waiting times are one example. Once again, I’m not creating difficulties here, but if we don’t take the time to really understand the quality of the data we are using, there can be no confidence in any claims we make on the basis of our results. If you want to know more about this, there has been some great work done by the Office for Statistics Regulation in its reports on the Quality Assurance of Administrative Data (QAAD).

**Value**

“Got it. Yes, that all makes sense. I’m on the home straight now. I’ve got the data, and everyone is happy. I’ve looked at it, and it looks like just what I need to answer my research question. I’m home and dry. No one else making things difficult for me.” Sorry to spoil your enthusiasm, but the hardest bit is yet to come.

One of the phrases I struggle with is “evidence-based decision making.” In my experience, reasonable people can reach different conclusions on the same evidence. What you decide depends on lots of things as well as the evidence. How you feel about something. Your values. What matters to you. Whenever I see a claim in the form of “the evidence says this, therefore you should do that,” I fear that the currency of research has been devalued. It prompts people like Michael Gove to say he has had enough of experts (who have been so wrong in the past).

The risk of making inflated claims is high, especially when your research is competing for interest with so many other stories of interest to your audience. Don’t be tempted. Don’t let projections masquerade as predictions. Correlation should not be mistaken for causation. Significant results (according to p values) are not a reason to act when effect sizes are low; with large administrative datasets and poor experimental design, it is easy to find a highly significant result of no research value whatsoever.

Think like a statistician when examining your results. Random results, especially outliers, are not necessarily interesting findings. Don’t present conclusions with too many significant digits and margins of error airbrushed out for dramatic effect. Avoid selective conclusions without context that lazily mislead. However strong the demand for certainty and the desire for the truth, we know that our findings are not conclusive—there are inevitably margins of error. Knowledge is always provisional.
These questions are not an extra difficulty but essential to good statistical practice. If we don’t ask them and insist on good answers, we undermine faith in the work of all our fellow practitioners.

Equally difficult is when we have findings that really do challenge received wisdom, that add to our understanding of the world, and that pose questions which really do need answers. In this situation, people really need to know. The researcher’s job is not finished when findings are published, only when those who need to understand them actually do so. And when there has been a proper debate and a decision on how to act. Who is our audience, and how do we connect with them? The researcher also needs to understand communication and see it as a skill essential to getting value from our results.

**Learning, Leadership, and Action**

So how can we proceed? As a community of researchers, we should be good at learning. Let us learn from what has gone before and learn from each other.

We can each be leaders in our own practice, in supporting our peers, and in not accepting poor practice. Together, we can lead the development of data as a critical part of national and global infrastructure, delivering on the citizen’s right to be informed. As leaders, we are making an investment that gives a rich return for better lives.

With so much lovely data and so many pressing research questions to be addressed, working together to tackle the difficult bits, the community of research can be at the heart of a transformation in the way we incentivize successful economies that deliver prosperity for all, a transformation in the way that society fights injustice and holds the powerful accountable, a transformation in the way we address climate change and create a sustainable future for our children and grandchildren, and a transformation in the way we develop stable and secure communities in turbulent times.

**About the Author**

John Pullinger is president of the International Association for Official Statistics and member of the Board of the Global Partnership for Sustainable Development Data. He was previously UK national statistician, president of the Royal Statistical Society, and chair of the United Nations Statistical Commission. He is a chartered statistician. He was also librarian and director general for Information Services at the UK House of Commons. He was made a Companion of the Order of the Bath (CB) in the Queen’s 2014 New Year’s honors list for services to Parliament and the Community. He is a graduate of the University of Exeter and of Harvard Business School. He has been awarded honorary degrees by the universities of Exeter, Essex, and the West of England. He is a visiting professor at Imperial College and fellow of the Academy of Social Sciences.