THE BIG PEEK

The data contained in tax returns, health and welfare records could be a gold mine for scientists — but only if they can protect people's privacy.

BY ERIKA CHECK HAYDEN

In 2011, six US economists tackled a question at the heart of education policy: how much does great teaching help children in the long run?

They started with the records of more than 11,500 Tennessee schoolchildren who, as part of an experiment in the 1980s, had been randomly assigned to high- and average-quality teachers between the ages of five and eight. Then they gauged the children's earnings as adults from federal tax returns filed in the

BARTHOLOMEW COOKE/TRUNK ARCHIVE

2000s. The analysis[1] showed that the benefits of a good early education last for decades: each year of better teaching in childhood boosted an individual's annual earnings by some 3.5% on average. Other data showed the same individuals besting their peers on measures such as university attendance, retirement savings, marriage rates and home ownership.

The economists' work was widely hailed in education-policy circles, and US President Barack Obama cited it in his 2012 State of the Union address when he called for more investment in teacher training.

But for many social scientists, the most impressive thing was that the authors had been able to examine US federal tax returns: a closely guarded data set that was then available to researchers only with tight restrictions. This has made the study an emblem for both the challenges and the enormous potential power of 'administrative data' — information collected during routine provision of services, including tax returns, records of welfare benefits, data on visits to doctors and hospitals, and criminal records. Unlike Internet searches, social-media posts and the rest of the digital trails that people establish in their daily lives, administrative data cover entire populations with minimal self-selection effects: in the US census, for example, everyone sampled is required by law to respond and tell the truth.

This puts administrative data sets at the frontier of social science, says John Friedman, an economist at Brown University in Providence, Rhode Island, and one of the lead authors of the education study[1]. "They allow researchers to not just get at old questions in a new way," he says, "but to come at problems that were completely impossible before."

PROBING THE POPULATION

In the past few years, administrative data have been used to investigate issues ranging from the side effects of vaccines[2] to the lasting impact of a child's neighbourhood on his or her ability to earn and prosper as an adult[3]. Proponents say that these rich information sources could greatly improve how governments measure the effectiveness of social programmes such as providing stipends to help families move to more resource-rich neighbourhoods.

But there is also concern that the rush to use these data could pose new threats to citizens' privacy. "The types of protections that we're used to thinking about have been based on the twin pillars of anonymity and informed consent, and neither of those hold in this new world," says Julia Lane, an economist at New York University. In 2013, for instance, researchers showed that they could uncover the identities of supposedly anonymous participants in a genetic study simply by cross-referencing their data with publicly available genealogical information (see *Nature* **497,** 172–174; 2013).

Many people are looking for ways to address these concerns without inhibiting research. Suggested solutions include policy measures, such as an international code of conduct for data privacy, and technical methods that allow the use of the data while protecting privacy. Crucially, notes Lane, although preserving privacy sometimes complicates researchers' lives, it is necessary to uphold the public trust that makes the work possible.

"IT SHOULD BE HARD TO GET ACCESS TO DATA, BUT IT'S VERY IMPORTANT THAT SUCH ACCESS BE MADE POSSIBLE."

"Difficulty in access is a feature, not a bug," she says. "It should be hard to get access to data, but it's very important that such access be made possible."

Many nations collect administrative data on a massive scale, but only a few, notably in northern Europe, have so far made it easy for researchers to use those data.

In Denmark, for instance, every newborn child is assigned a unique identification number that tracks his or her lifelong interactions with the country's free health-care system and almost every other government service. In 2002, researchers used data gathered through this identification system to retrospectively analyse the vaccination and health status of almost every child born in the country from 1991 to 1998 — 537,000 in all. At the time, it was the largest study ever to disprove[2] the now-debunked link between measles vaccination and autism.

Other countries have begun to catch up. In 2012, for instance, Britain launched the unified UK Data Service to facilitate research access to data from the country's census and other surveys. A year later, the service added a new Administrative Data Research Network, which has centres in England, Scotland, Northern Ireland and Wales to provide secure environments for researchers to access anonymized administrative data.

In the United States, the Census Bureau has been expanding its network of Research Data Centers, which currently includes 19 sites around the country at which researchers with the appropriate permissions can access confidential data from the bureau itself, as well as from other agencies. "We're trying to explore all the available ways that we can expand access to these rich data sets," says Ron Jarmin, the bureau's assistant director for research and methodology.

In January, a group of federal agencies, foundations and universities created the Institute for Research on Innovation and Science at the University of Michigan in Ann Arbor to combine university and government data and measure the impact of research spending on economic outcomes. And in July, the US House of Representatives passed a bipartisan bill to study whether the federal government should provide a central clearing house of statistical administrative data.

Yet vast swathes of administrative data are still inaccessible, says George Alter, director of the Inter-university Consortium for Political and Social Research based at the University of Michigan, which serves as a data repository for approximately 760 institutions. "Health systems, social-welfare systems, financial transactions, business records — those things are just not available in most cases because of privacy concerns," says Alter. "This is a big drag on research."

UNSOUGHT INTIMACY

Feeding those concerns is the rising public unease about online privacy in general. Private companies known as data brokers operate on a vast scale, collecting and selling information about Internet searches, online purchases and other data streams that can be combined to draw surprisingly intimate conclusions. In one famous example, the US retailer Target inferred that a teenage girl was pregnant based on her purchases there, and it began sending her coupons for baby products; her father was alerted to his impending grandchild only when the coupons arrived at the family's home. In a 2014 study[4] of data brokers, the US Federal Trade Commission pointed out the many ways in which this kind of information could harm consumers. People who buy products such as blood-sugar monitors, for instance, might be placed into a 'diabetes risk' marketing category that could be used by an insurance company to pinpoint a potential customer as high risk.

Many researchers argue, however, that there are legitimate scientific uses for such data (see *Nature* **488,** 448–450; 2012). Jarmin says that the Census Bureau is exploring the use of data from credit-card companies to monitor economic activity. And researchers funded by the US National Science Foundation are studying how to use public Twitter posts to keep track of trends in phenomena such as unemployment.

But not everyone makes the distinction between commerce and academia, says Lane. "People conflate the concern about big data being used for private-sector purposes to make money with big data being used for research." In March 2014, for instance, while aiming to

significantly boost consumer privacy through a new data-protection regulation, the European Parliament proposed limiting the use of personal health data for research without specific consent, which would have severely curtailed researchers' access to those data. After objections from organizations such as the London-based biomedical-research charity the Wellcome Trust, the proposal looks likely to be jettisoned, but its fate will not become clear until 2016, when the final text of the regulation comes up for approval.

One solution to the privacy concerns has been to keep data under lock and key, tightly restricting who can access it. At the US research data centres, for instance, investigators are not allowed to take smartphones or flash drives into the rooms where they will use the centre's computer terminals. The computers themselves contain no data, but only link remotely to secure servers.

TECHNICAL ANSWERS

Computer scientists and cryptographers are experimenting with technological solutions. One, called differential privacy, adds a small amount of distortion to a data set, so that querying the data gives a roughly accurate result without revealing the identity of the individuals involved. The US Census Bureau uses this approach for its OnTheMap project, which tracks workers' daily commutes. Researchers at the bureau use actual data to build a statistical model based on where individual workers commute each day. They then build a synthetic data set that fits the model, but does not contain the actual data. This synthetic data set is released to the public, allowing users to draw accurate conclusions about transport and economic trends without tracking the exact movements of real individuals. Researchers are still learning to trust synthetic data, however, so few papers that have been published on this subject go beyond demonstrating the methods.

In any case, although synthetic data potentially solve the privacy problem, there are some research applications that cannot tolerate any noise in the data. A good example is the work showing the effect of neighbourhood on earning potential[3], which was carried out by Raj Chetty, an economist at Harvard University in Cambridge, Massachusetts. Chetty needed to track specific individuals to show that the areas in which children live their early lives correlate with their ability to earn more or less than their parents. In subsequent studies[5], Chetty and his colleagues showed that moving children from resource-poor to resource-rich neighbourhoods can boost their earnings in adulthood, proving a causal link.

Secure multiparty computation is a technique that attempts to address this issue by allowing multiple data holders to analyse parts of the total data set, without revealing the underlying data to each other. Only the results of the analyses are shared.

For instance, in 2010, the US Defense Advanced Research Projects Agency (DARPA) asked a team of cryptographers to develop a secure multiparty computation protocol to analyse the paths of commercial satellites and head off costly collisions. Currently, companies do this by sharing their orbit data, which they consider proprietary, to a trusted third party that performs the analysis. But DARPA concluded that secure multiparty computation could be used to predict possible collisions just as effectively, albeit a little more slowly.

"THE LESSON IS NOT TO UNDERESTIMATE PUBLIC CONCERNS. PUBLIC TRUST IS VERY FRAGILE."

In 2015, the Estonian company Cybernetica, based in Tallinn, said that it had used similar techniques to analyse financial filings of companies to detect tax fraud. It is also jointly analysing records from the country's tax and education ministries to explore whether university students who hold jobs fail their courses more often than those who focus exclusively on their studies.

There are still some problems in need of technical solutions — especially as government agencies look beyond their own walls. For instance, the Census Bureau wants to combine its internal data on the formation and activities of companies with public data on patents to examine the factors that drive corporate innovation. But it could be relatively easy to unmask the identities of companies included in the analysis by matching them to information in the public patent database. Jarmin's team has not yet worked out an approach that adequately protects privacy.

But for the most part, technical solutions are now being put in place. Increasingly, what looks likely to hold up the research is a lack of clear ethical and legal guidance about how data on individuals can be used — for all purposes, including research.

Pam Dixon, executive director of the World Privacy Forum in San Diego, California, points to programmes such as India's national identification-card system, launched in 2010. This effort provided more than 900 million people with biometric identity cards that were linked to photographs, fingerprints and iris scans. The cards were supposed to be voluntary, and were used to identify rightful recipients of social benefits such as fuel and unemployment aid.

But the country did not create a legislative framework to govern the use of the cards. They were soon discussed as gateways for a variety of essential services, such as salary payment and marriage registrations. This violated the original spirit of the programme, critics contended, because data from the cards was not supposed to be coerced from individuals. The Indian Supreme Court ruled such uses of the system illegal on 11 August, but the country's Parliament has still not enacted a governing framework.

Likewise, in 2013, the United Kingdom launched the care.data programme to link records from patients' visits to general practitioners with their records from other parts of the health-care system, but there was no clear guidance on how the project's data were to be used. After it was revealed that the database designed to distribute patient data had inappropriately released some information to private entities — such as actuaries, which aid insurers in setting insurance rates — care.data came under fire. On 2 September, the National Health Service (NHS) said that the government will conduct a review of the security of NHS data and develop new opt-out and consent provisions. The system is intended to be available to all patients by 2016.

In the meantime, says Nicola Perrin, head of policy at the Wellcome Trust, the fallout has created huge delays in existing research projects, including clinical trials and health evaluation, audit and service research. Researchers in charge of SABRE, a large cohort study examining how diabetes and heart disease affect people of different ethnicities, have not received patient updates since March 2014; as a result, they risk sending requests for information to families whose loved ones may have died. The episode serves, for Perrin, as a cautionary tale about how the power of data could backfire if social unease with its uses is not addressed as soon as possible. "The lesson is to not underestimate public concerns," she says. "Public trust is very fragile — it's difficult to build and easy to break." ■

Erika Check Hayden *is a reporter for* Nature *in San Francisco, California.*

1. Chetty, R. *et al. Q. J. Econ.* **126,** 1593–1660 (2011).
2. Madsen, K. M. *et al. N. Engl. J. Med.* **347,** 1477–1482 (2002).
3. Chetty, R., Hendren, N., Kline, P. & Saez, E. *Q. J. Econ.* **129,** 1553–1623 (2014).
4. Ramirez, E., Brill, J., Ohlhausen, M. K., Wright, J. D. & McSweeny, T. *Data Brokers: A Call for Transparency and Accountability* (Federal Trade Commission, 2014).
5. Chetty, R., Hendren, N. & Katz, L. F. NBER Working Paper No. 21156 (2015); available at http://www.nber.org/papers/w21156