Harnessing new social data for effective social policy and service delivery

Workshop background note

OECD

TE TAI ŌHANGA
THE TREASURY

Australian Government
Australian Institute of Health and Welfare

accenture

Introduction

Better data can improve the ability of social protection systems to identify and address people's needs. Different risks such as job loss, income decline, health problems and family breakup interact in complex and changing ways. Similarly, the manner in which social assistance, social insurance and social care – the social protection system broadly understood – address these risks evolves over time. Yet while many risks influence each other, policy responses are frequently fragmented, addressing only one problem at a time. The failure to prevent risks from materialising in the first place or to mitigate their consequences quickly can escalate costs across the social policy spectrum. Smart use of data can contribute to a more effective social policy that makes it less likely that such a situation arises.

In acknowledgement of this reality, the OECD ministers and representatives responsible for social policy called for the establishment of a *Social Data for Tomorrow* programme in their Ministerial Policy Statement of May 15 2018. The workshop on *Harnessing social data for effective social policy and service delivery* aims to lay the groundwork for a programme supporting member countries' activities using new data developments to enhance social policy. The workshop is organised with support of the Treasury New Zealand, the Australian Institute of Health and Welfare and Accenture. Representatives from statistical offices and other government agencies and researchers will present examples of how they leverage social data innovations to improve social policy design, implementation and evaluation. A thematic focus will illustrate how different actors draw on new social data approaches to improve child well-being.

This background note introduces the topics that will be discussed during the workshop. It first describes innovations in the realms of social data collection, of data linking and of utilising data from non-traditional sources. Second, it outlines novel ways in which social policy planning, implementation and evaluation make use of these data.

Data innovations

Recent advances in data collection and linking of different data sources have an enormous potential for better evidence-based policymaking. Improved access to administrative data and tools such as web-based surveys already offer powerful new ways of identifying the needs of different population groups and following their pathways through multiple systems. In addition, the digital recording of people's online and real life activities, including how they interact with public services, generates data of previously unimaginable volume in a variety of formats. Such big data are not always representative, reducing their accuracy (veracity) compared to traditional data sources. But big data's more immediate availability (velocity) increases their comparative value, in particular since falling survey response rates are creating quality and cost challenges. More rapid evaluations of social policy interventions based on real-time data from multiple sources could greatly benefit social outcomes.

Innovations in data collection

Statistical agencies have always adjusted the topic coverage of the data they collect and disseminate. Nevertheless, in recent years, their surveys have been covering a broader range of social risks and needs. For example, in 2009, the EU Statistics on Income and Living Conditions (EU-SILC) first included a material deprivation module. This module is now the basis for a 'material and social deprivation' indicator and its most recent iteration contains optional questions on housing instability. The 2011-2013 Australian Health Survey included anthropometric, biomedical and environmental measures, following the example of the Health Survey for England that collects biomedical measures for a sub-sample of respondents. In some cases, the expansion covers topics that were previously predominantly explored through qualitative research. For instance, the Mexican National Institute for Statistics and Geography carries out a slew of surveys related to perceptions of safety and experiences of crime, including a survey on intimate partner violence. The New Zealand General Social Survey collects information on social connection and social networks.

Many general and specialised surveys are putting more effort into collecting data on subjective perceptions of social risks and needs. At the OECD, the 2018 *Risk that Matters* survey collected representative data for 21 countries on people's perceptions of the main social and economic risks they faced and of the government's ability to address those risks. Planned further waves will allow comparing risk perceptions over time. The *Trustlab* project explored trust in institutions through an online behavioural game with a representative sample of 1 000 respondents each in six OECD countries.

Expanded survey activities do not only cover a broader range of topics, but also a more comprehensive share of the population. While population censuses typically cover a country's entire population, household surveys usually only cover the population living in households. The institutionalised population and other individuals not living in households – many of whom are among the most vulnerable members of society – are excluded. But different national statistical offices are trying to address this shortcoming. The United Kingdom's Office for National Statistics has produced a position paper and commissioned a report on including non-household populations in data collection and in estimates of personal well-being and destitution. The French National Institute for Statistics and Economic Studies carried out surveys of the homeless in 2001 and 2012. Moreover, the Institute's 2009 and 2019 longitudinal surveys on the integration of newcomers (ELIPA and ELIPA2) collected information on recent immigrants who had signed an integration contract, including on people living in collective housing.

For cost and convenience reasons, survey data collection is increasingly moving online. Some online surveys use probability based sampling. They typically recruit respondents offline and can provide computer and internet access to those that do not already have it. Other surveys have non-probability sampling and recruit their participants online. Since the share of 16-74 year olds that have used the internet in the last three months is still below 90% in two thirds of OECD countries with recently available data, and survey volunteers are not representative of internet users, some worry that online surveys with non-probability sampling are inherently non-representative. However, survey firms aim to ensure representativeness through various means, such as pre-selecting interviewees based on their demographic characteristics and weighting. Online surveys can allow the implementation of more complex survey instruments. The previously mentioned Trustlab is one example of the combination of basic survey questions with behavioural games. Previously, such experiments were usually confined to university labs, where students – a group highly unrepresentative of the population at large – were the main participants. Online surveys can also allow more frequent follow-ups, for example through text messages. Potentially, information that smartphones collect passively can also enhance surveys. Of course, this requires that affected respondents are fully informed and consent to the additional data collection. Moreover, the organisations that gather and analyse such combined data need to have even more stringent data protection mechanisms in order to minimise the chances of privacy breaches.

Statistical offices are increasingly augmenting or even replacing survey with administrative data. This can increase response accuracy and lower the time costs that surveys impose on respondents. For example, in a number of European countries, income and some demographic information for the EU Statistics on Income and Living Conditions is gathered through registers rather than survey questions. Portugal's National Statistical Office studied the feasibility of replacing questions in the 2021 population and housing census with administrative data-based measurements. They identified 12 out of the 27 mandatory variables for which administrative data can be an acceptable replacement in the medium to long term.

Innovations in administrative data linking

Linked administrative-survey data can facilitate deeper insights into people's long-term pathways. Data linkage can occur on an ad-hoc basis, provided the organisations holding the data agree and survey respondents consent. However, an increasing number of research organisations create linked files on a regular basis. Examples include the German PASS-ADIAB dataset that links data from a labour market and social protection survey to administrative labour market biographies; the US HRS-SSA dataset that combines Health and Retirement Survey and Social Security Administration records; and the Spanish Labour Force Survey that is augmented with wage information from the social security and tax databases.

Given that merged administrative-survey files contain more detailed and sensitive data, access conditions are more stringent than for survey data alone. In some cases, the data are completely non-public. In others, researchers have to go to an on-site safe-room at the data-holding institution, or access them remotely from a safe room at a different institution. Whatever the mode of access, the requirements placed on the researchers, including their legal liability for negligent or deliberate data mishandling that threatens respondents' anonymity, are higher than for regular survey public use files.

In many countries, survey participants have to actively consent to their responses being linked to administrative records. The share that does so varies strongly. A low share can severely limit the usefulness of the linked data: The resulting sample may be unrepresentative of the target population (the so-called consent bias) (Sakshaug et al., 2012[1]) or contain too few observations to answer relevant research questions. Consent rates may be higher in face-to-face or telephone interviews compared to web surveys (Thornby et al., 2018[2]).

Administrative databases can also be linked with each other. This linking, which may involve data from different government departments or levels, can reveal how life risks are related and what assistance people receive from different parts of the government. In the ideal case, it can also contribute to breaking down topic silos and to improving cooperation across departments. Government agencies can use the resulting merged database to guide, monitor and evaluate evidence-based social policymaking. Furthermore, by opening up the data to external researchers, they can widen the knowledge base on people's risks and needs and on the effectiveness of social policy interventions.

To make it easier for researchers to gain approval for using linked data, a number of OECD countries have created framework institutions that facilitate data merging. Among these are Canada's Social Data Linkage Environment, the United Kingdom's Administrative Data Research Partnership and the United States' Census Bureau's Data Linkage Infrastructure. Details differ, but these programmes typically combine several functions such as creating an inventory of data sources, reviewing research proposals, helping researchers gain approval from concerned agencies, linking the data and providing secure access.

The linkage programmes can help overcome the practical and legal hurdles to data access for researchers. However, they do not necessarily serve the needs of national, regional or local government agencies wishing to access data held at another agency to continuously support or monitor the implementation of social policies. The legal requirements for data sharing vary between jurisdictions, agencies and policy areas. While agreements that outline the sharing agencies' responsibilities and allowed data uses need to be tailored, creating standardised data sharing processes and model agreements can facilitate inter-agency data sharing.

A challenge arises when different databases do not have consistent personal identification numbers. In countries that have universal national personal identification numbers, this is unlikely to be a problem. However, even there, records belonging to the same person may fail to be connected because of data entry mistakes. In other countries, identification numbers differ across policy domains. In these cases, linking has to rely on imperfect identifiers such as names and dates-of-birth. Record linkage based on imperfect identifiers can be deterministic (using pre-determined rules to classify records as belonging to the same person) or probabilistic (assigning match weights based on how much the imperfect identifiers correspond to each other and linking records whose match weight exceeds a certain threshold). Deterministic linkage can lead to few errors but too few matches; while probabilistic linkage creates more links, including some between records belonging to different people (Harron et al., 2017[3]).

Different procedures can preserve individuals' privacy while linking databases. One possibility is to transfer the matching process to a third party. This third party only accesses the identity information to create cross-database linkage keys, but does not receive the remainder of the datasets. As an additional safety measure, the remaining data may be altered to make it less likely that individuals can be identified based on their records. This can include leaving out finer-grained geographic details, aggregating continuous into categorical data (data masking) or adding noise to the data (data perturbation). The 2015 OECD Health Policy Study on Health Data Governance – Privacy, Monitoring and Research describes examples of health data de-identification within broader data governance frameworks from several countries.

Public engagement work on data uses and linkage, as for instance undertaken by Australia and the United Kingdom, can have multiple benefits. First, it can highlight the potential benefits that can flow from data merging and analysis, including cost savings from replacing survey questions with administrative data and from a more holistic social policy, while also describing privacy and data security safeguarding mechanisms. Improved public communication may contribute to higher consent rates for data linkage, though this is far from certain. Second, it can allow data governance frameworks to reflect public attitudes on privacy and data security. In Australia, ongoing public consultations on the Data Sharing and Release legislation for example revealed that the community was sceptical of using public data to 'target' individuals to assess their eligibility for a government programme or service. Accordingly, the proposed legislation does not authorise such uses (Australian Government Department of the Prime Minister and Cabinet, 2019[4]).

Innovations in using non-traditional data sources

The ever-growing digital footprint of OECD citizens creates new sources of data for social indicators and social policy research. Some of these data are structured, meaning that they either are already stored in databases with rows and columns or can be easily transferred into one. Structured data include records from credit card companies, credit agencies and hospitals. Other data are semi- or unstructured and hence require substantial efforts to be transformed into a shape that can be analysed. These can come from the ever-increasing number of sensors that record individuals' locations, workout activities and sleep patterns; social media posts and internet searches; pictures and videos; satellite data, and others. For example, people may vent about the costs of childcare in social media posts. These posts could form the basis of an indicator on the perceived affordability of childcare. However, to extract the relevant information, analysts first need to identify key words associated with childcare affordability and then scan social media posts for these terms.

Some statistical offices and other government agencies are making use of these data to create indicators. These statistics, most of which are still experimental, can cover new topics, or be more cost efficient or frequent than statistics based on survey data. Most existing applications are in policy realms outside of social policy. For example, as a first official statistic constructed from 'big data' worldwide, Statistics Netherlands created a traffic intensity measure based on cars passing sensors on motorways. A research consortium that included Statistics Estonia tested the use of phone GPS data to measure tourism. Researchers have estimated inflation based on web-based prices. The U.S. Bureau of Labor Statistics has started incorporating prices directly reported from a retailer into its consumer price index, to complement the pricing information its own enumerators assemble in sampled stores and on websites.

Academic researchers are equally exploring ways to use 'big data' to understand social risks and needs. For example, Italian researchers suggested that more geographically disaggregated poverty measures could be created based on combining official statistics with mobile phone data (Marchetti et al., 2015[5]). Google searches have been used to estimate indices of job search intensities (Baker and Fradkin, 2017[6]) and well-being (Algani et al., 2016[7]). Social media posts could be the basis for a consumer confidence indicator (Daas and Puts, 2014[8]); and Facebook advertising platform and geo-tagged Twitter data have been used to estimate immigrant stocks (Zagheni, Weber and Gummadi, 2017[9]) and study the link between short-term mobility and long-term migration (Fiorio et al., 2017[10]), respectively.

A first challenge for the use of data from alternative sources is gaining access and creating a legal framework. This is not an issue for web scraped data, where permissible, but applies for any proprietary data. Companies may be unwilling to provide outsiders data access. When they agree, contracts needs to define access conditions, including whether the data are transferred and who is authorised to analyse them, as well as payment and liability details. A 2017 OECD Statistics Working Paper discusses this in detail (Klein and Verhulst, 2017[11]). Whenever the data-providing firms could gain undue advantages over their competitors from advance knowledge of resulting indicators – as may be the case for unemployment and other economic indicators that often lead to market movements – safeguards need to be put in place.

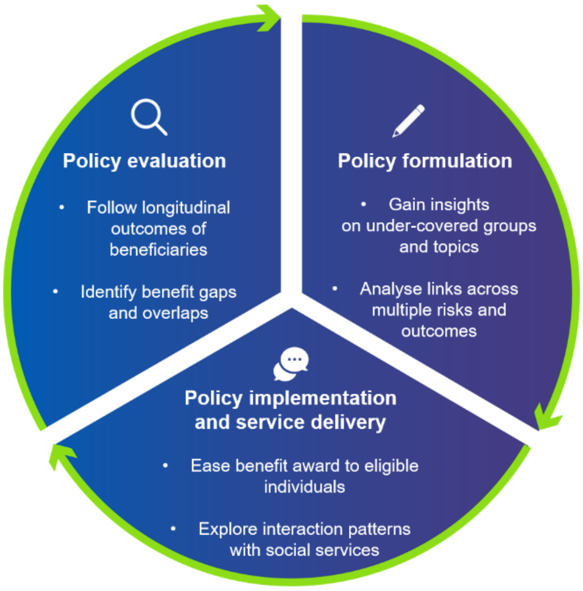
A second concern is the accuracy and representativeness of the underlying data or resulting statistic. Google Flu Trends is a well-known cautionary tale (Lazer et al., 2014[12]). The indicator is aimed at predicting flu trends based on people searching for information on flu-related symptoms online. The indicator was initially very accurate, but started to fail after a few years. One of the reasons was that people's online search behaviour changed, in part due to increased media reports on the flu. While this issue is related to the analytical strategy, another threat is that the compiled data may not be representative of the population whose outcomes the researchers or statisticians would like to study. Older, rural and more socio-economically disadvantaged people may be less likely to be have constant internet access, use a credit card or drive on toll roads. Statisticians may be able to adjust for the lack of representativeness, up to a degree. But these adjustments are necessarily harder and require more assumptions than creating analytical weights for surveys with probability-based sampling.

A third problematic issue is the potential erosion of individuals' privacy. The Council Recommendation concerning Guidelines covering the Protection of Privacy and Transborder Flows of Personal Data [C(80)58/FINAL] recognise this tension, noting that "more extensive and innovative uses of personal data bring greater economic and social benefits, but also increase privacy risks". Depending on the nature of the data, companies might have to seek explicit consent from concerned individuals before making data available for analysis by government agencies. This can be a safety measure against privacy intrusions, although only if consent is actually optional (i.e. individuals can use or acquire the companies' goods or services even when not giving consent) and informed (i.e. the consent clause is not buried in lengthy and hard-to-understand terms of service). It can lead to the same 'consent bias' that was mentioned in the administrative data linkage section. While the previously discussed uses of non-traditional data are stand-alone projects without data linkage, combining private and public microdata could potentially lead to transparent citizens whose every step, internet search, e-mail, social media post, purchase and interaction with a government agency are recorded and analysed. Although still in its early stages and with uneven local implementation, China's Social Credit system, which combines public and private data and is intended to steer citizens towards socially desired behaviour through incentives and punishment, is a step in this direction (Ohlberg, Ahmed and Lang, 2017[13]).

Applications of data-related innovations in social policy

The prior section shows that data innovations in the social policy sphere are often still in their infancy. Governments that want to strengthen these innovations need to invest substantial resources. Moreover, they have to grapple with thorny issues such as what limits are required to preserve individual privacy and how to update security procedures continuously to prevent data breaches. While the topic is complex, examples of public engagement efforts show that these can help governments make decisions about how much to invest and which limits respect predominant views about data privacy and security. Governments considering social data investments have to take into account not only the costs but also the potential value. Creating new ways to measure and analyse citizens' social and economic outcomes can benefit all stages of the policy making cycle (see Figure 1). The following sections present examples of applied uses of new social data in these different stages. In addition, it highlights additional concerns that may emerge in particular applications.

Figure 1: Examples of uses of new social data in different stages of the social policy making cycle


Policy evaluation
• Follow longitudinal outcomes of beneficiaries
• Identify benefit gaps and overlaps
Policy formulation
• Gain insights on under-covered groups and topics
• Analyse links across multiple risks and outcomes
Policy implementation and service delivery
• Ease benefit award to eligible individuals
• Explore interaction patterns with social services

Policy formulation and design

Compared to the phases of policy implementation and evaluation, it could be argued that data innovations will shift the way in which policies are designed in a less fundamental manner. Nevertheless, a more comprehensive and up-to-date understanding of the social needs of the entire population and of the links between different social policy areas can allow governments to identify policy misalignments and to create more effective and less expensive evidence-based policy interventions.

New social data can lead to a more complete understanding of the outcomes of population groups and of specific risks about which little was previously known. For examples, a survey of the homeless population in France revealed that a quarter among them had previously been in foster care (Yaouancq and Duée, 2014[14]). This finding influenced thoughts about providing additional support for 'care leavers' about to turn eighteen. The Mexican 2014-2018 Programme to prevent, address, punish and eradicate violence against women makes ample reference to statistics on intimate partner violence based on specialised surveys. Increased access to different administrative databases, often combined with new data visualisation techniques, has deepened the understanding of policy makers and the public alike on how much life chances vary between adjacent neighbourhoods (Chetty et al., 2018[15]). Information from unstructured data sources may also be helpful to understand such geographic variations not captured by survey data: A study from Ontario analysed call records to a human services helplines to reveal regional variations in service gaps (Dillon Consulting, 2018[16]).

Combined administrative data allows detailed insights into the services individuals access jointly and into their outcomes. Ultimately, this could make it possible to identify where investments in one policy area can improve outcomes in another. For instance, a Scottish study analysed the health outcomes and health services usage of the homeless and near-homeless (Waugh et al., 2018[17]). Similar research by the Australian Institute for Health and Welfare combined data from different agencies to understand the characteristics and housing outcomes of young people in contact with the youth justice or child protection systems (Australian Institute of Health and Welfare, 2016[18]).

The potentially biggest change from new social data for the policy planning process, however, might be the possibility to create social indicators at more frequent intervals. Unemployment rates are typically released monthly while poverty rates are released yearly, creating the potential for more reactive employment compared to social policies. Yet buried in administrative data or data from alternative sources is the potential to create more ‘instant’ indicators. These ‘now-casts’ do not necessarily need to have the same statistical validity as official statistics do. Rather, policy makers could use them to quickly observe changes in social outcomes, allowing them to start thinking about potential ways to address problems that are on the cusp of worsening. As of now, published social policy-relevant applications appear rare.

Policy implementation and service delivery

The application of social data innovations could drastically alter the delivery of social assistance and social services. Agencies have started using linked data to automatically grant benefits, segment their clients and target interventions, even if many experiments are still in their infancy. But precisely because they seek to analyse and act on individual data records, rather than to understand patterns in the population, policy makers need to be mindful of the threats to privacy, accountability and non-discrimination these applications in particular could pose.

The combination of data from different sources can allow identifying individuals who are eligible for social benefits and services. For example, the Chilean Social Registry of Households combines self-reported and administrative data from various government agencies, and provides this information to 27 institutions and 345 municipalities. Based on these data, a socioeconomic qualification score places households in income and socioeconomic vulnerability deciles. Some services use this score to assign benefits directly. Others employ it as an input for their own selection instrument or to prioritise potential beneficiaries (Ministerio de Desarrollo Social, 2018[19]). The Belgian Crossroads Bank for Social Security coordinates information exchanges between the country’s 3 000 social security actors, allowing the automatic granting of several benefits.

In other contexts, it could be possible to identify potential beneficiaries and encourage them to apply, but published examples are lacking. An exception is a small initiative in North Carolina: There, two providers of human and welfare services share relevant data and encourage clients who did not apply for all available benefits to do so (Epstein and Maxwell, 2016[20]). Creating tailored ‘service packages’ based on individual data-based profiling is also a theoretical possibility, but a published account of such a practice could not be identified.

Other possible applications aim to identify high-risk individuals among existing clients of a social service agency or in the population at large. For instance, the New Zealand Work and Income Youth Service has a risk-scoring algorithm that aims to predict which school leavers are at high risk of becoming long-term unemployed. The algorithm considers factors such as how the former student did at school, whether their parents received welfare benefits and whether they were in contact with child protective services. Service providers then approach those deemed most at risk (Stats NZ, 2018[21]). A number of child protective agencies in different countries are thinking about using available information to identify children potentially at risk. For instance, the Allegheny Family Screening Tool combines information from past calls to the agency with data from the justice system, psychiatric and drug treatment services and welfare offices. When the agency receives a call about a child or family, the tool assigns a risk score. A high score triggers further investigation and a caseworker visit.

The first major concern with using data analysis for individually targeted interventions is the connected invasion of privacy. Not only do involved agencies gather and link sometimes extremely sensitive information about individuals, they then use them to study individual cases and to potentially reach out to individuals. This may include situations where the attention may not be welcome, such as when child protection agencies contact parents. Views on how problematic the entailing reduction in privacy is vary from person to person and country to country. For example, while the Allegheny Family Screening Tool was implemented and garnered interest from other cities and states in the United States, a similar planned intervention in the Danish municipality of Gladsaxe was blocked because it was in conflict with privacy legislation. The assessments may also depend on the social service in question. People may for example be more tolerant of personalised predictions when they intend to protect vulnerable children, but less so in other instances.

A second concern is the process's transparency and accountability. When algorithms support service providers in taking decisions, and even more so when decision making is fully automated, it can be difficult for case workers and affected citizens to understand how decisions come about. Quite aside from the opacity of algorithms themselves, there may be mistakes in the coding and in the underlying data since no data entry system is perfect. Such errors can make someone seem ineligible for a benefit they are legally entitled to, and vice versa. Virginia Eubanks (2018[22]) offers an eloquent account of the effects the automation of welfare award decisions had on affected families that suddenly – and sometimes mistakenly - lost their benefits. The Recommendation on Artificial Intelligence (AI) adopted by the OECD Council at Ministerial level on 22 May 2019 (OECD/Legal/0449) recommends that those that are adversely affected by an AI system should be able to "challenge its outcomes based on (…) easy to understand information on the factors (…) that served as the basis for the prediction, recommendation or decision." This suggestion implies that public agencies either need to develop algorithms in-house or that they need to be able to fully access and understand algorithms developed by external parties, including by private companies. Citizens need to be able to appeal decisions; and judges or other arbitrators should receive training on the fallibility of algorithms (Citron, 2007[23]). External reviews could evaluate the quality of algorithms according to how accurate, cost-effective and fair they are.

A third concern is discrimination. At first glance, this may seem paradoxical: while human decision-makers have to battle conscious and unconscious biases and have difficulties gathering and keeping all relevant pieces of information in mind, an algorithm can consider hundreds of variables at a time and come to an 'objective' decision. But evaluations of algorithmic decisions have found that they can be discriminatory in practice even when the variables along which discrimination is measured, such as gender, ethnicity or age, are not themselves included. As Osoba and Walker (2017, p. 17[24]) state, "applying procedurally correct algorithm to biased data is a good way to teach artificial agents to imitate whatever bias the data contains". Data may be biased for different reasons. First, certain population groups could be over- or under-represented in the databases. For example, child abuse may be equally common in poor and rich households, but neighbours may be more likely to notice and report it when it happens in families living in apartment buildings with thin walls than in families living in stand-alone homes. Second, the algorithm might mirror decisions taken by biased individuals. This appears to have for example happened in applications in the criminal justice system. Moreover, algorithmic decision support can create a self-reinforcing feedback loop. By pointing authorities to pay more attention to certain population groups (or, in the case of policing, areas), more data is gathered that may provide more evidence that even more attention should be focused on the concerned group (O'Neil, 2016[25]). Researchers are working on tweaks that prevent algorithms from learning biases, although there may be a certain trade-off between accuracy and non-discrimination.

Policy evaluation

Once policies and programmes are in place, evaluations can assess their intended and unintended effects. Knowledge from these evaluations can then feed back into the policy making process, laying one of the foundations for evidence-based policy making. While policy and programme evaluations are nothing new, the novel data context allows for several innovations. These include longer-term and ex-post evaluations due to improved administrative data access and more immediate evaluations thanks to easier access to big data. A broader data base however does not solve all difficulties that evaluations encounter. Finding the 'right' counterfactual - a prerequisite to evaluate the causal impacts of an intervention - outside of randomised control trials remains challenging despite important methodological advances in quasi-experimental methods in the past few decades.

Evaluations using administrative data can measure different impacts over long periods. Prior to the easier – but often still difficult – access to administrative microdata, evaluations had to be planned before a policy or programme was implemented. The affected population group and, in the case of a counter-factual evaluation, the control group had to be surveyed. Survey dropout was a serious threat to the evaluation, and the topic and time range of measured outcomes were often limited. The increased availability of administrative data can address some of these shortcomings. As long as it is for example possible to link programme recipients and non-recipients to their administrative records, some of their outcomes can be measured even if they stop responding to surveys. Potentially, such evaluations can also reveal side effects more easily.

Administrative data can provide more accurate measures of key programme outcomes. For instance, a recent evaluation of anti-poverty programmes based on both survey and administrative data found that survey data often under-stated the incomes of low-income respondents. As a result, evaluations based on survey data alone found a lower poverty-reducing impact of the studied programmes (Meyer and Mittag, 2019[26]). Moreover, potentially eligible beneficiaries for a social benefit can be identified. A list of such individuals can then serve as a sampling frame for studies on why they do not apply and on policy interventions that might incentivise them to do so. Examples include an evaluation of policies to increase applications for food stamps among likely eligible individuals (Finkelstein and Notowidigdo, 2019[27]) and the TAKE project that applies different research methodologies including microsimulation models, a field experiment and microeconometric analyses based on survey and administrative data to study non-take up of various benefits in Belgium (TAKE-Project, n.d.[28]).

Another major gain from data innovations for social policy evaluations can be more immediate evaluations. More routine access to administrative data can already allow researchers to design evaluations in a detailed manner prior to a policy change or programme introduction and to start carrying them out and having first results shortly after (Langedijk, Vollbracht and Paruolo, 2019[29]). Short computer- and mobile-phone based surveys can provide complementary data at more frequent intervals. And finally, ‘big’ data from non-traditional sources can deliver further information that is available more quickly or that is complementary, for instance allowing insights on public attitudes towards a policy (Global Pulse, 2016[30]).

Conclusion

The discussion above shows that many OECD countries are moving in similar directions regarding data innovations. Many statistical offices are expanding surveys to cover more topics relevant for social policy. Countries are also increasingly linking administrative databases amongst each other and with survey data and creating specific entities tasked with this linking. The existence of different approaches to survey coverage and data linking offers scope for mutual learning between OECD countries. In contrast, few academic researchers and even fewer government agencies are so far using data from non-governmental sources to understand and shape social policy. It would be desirable to reflect systematically on the possibilities these data could offer but also on the potential privacy implications.

‘New uses’ of data innovations are less common in social policy than in other policy areas. In many countries, sub-national government agencies, including individual city administrations, and researchers propose applications more frequently than national governments do. Given the repercussions such applications may have for social policy design and implementation but also for individuals’ privacy and administrative accountability, the current moment presents an ideal point in time to ponder a number of important questions.

A first area of key questions concerns the scope and limits of different data uses in social policy design, evaluation and implementation and service delivery. For example, should any administrative database containing personal data that is held by a government agency be available for linking with other databases? To what extent should statistical offices and other agencies try to obtain data from private companies, and should these data ever be linked to public data about individuals? Should citizens consent explicitly before their survey responses are linked to administrative records, or should their consent be implicitly assumed as long as they do not formally withdraw it? Which categories of stakeholders should have permission to access these data– civil servants developing implementation plans for new social policy interventions; front-line social services staff working with individual clients; researchers studying the impacts of a programme? And how much can decisions about the targeting of benefits and interventions be based on evidence from linked databases, and how much can these decisions be automated? At a practical level, a series of operational questions arise: How should the reached consensus be translated into appropriate laws and regulations? How can governments, academic researchers and companies best cooperate in utilising data from 'alternative' sources? And what is the best and most cost-effective way to create or access the necessary technological infrastructure and to build the needed analytical capacities?

References

Algani, Y. et al. (2016), "Big Data Measures of Well-Being: Evidence From a Google Well-Being Index in the United States"*, OECD Statistics Working Papers*, No. 3, OECD Publishing, Paris, https://doi.org/10.1787/5jlz9hpg0rd1-en. [7]

Australian Government Department of the Prime Minister and Cabinet (2019), *Data Sharing and Release Legislative Reform Discussion Paper*, https://www.datacommissioner.gov.au/resources/discussion-paper. [4]

Australian Institute of Health and Welfare (2016), *Vulnerable young people: interactions across homelessness, youth justice and child protection*, AIHW, Canberra. [18]

Baker, S. and A. Fradkin (2017), "The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data", *The Review of Economics and Statistics*, doi: 10.1162/REST_a_00674, pp. 756-768, http://dx.doi.org/10.1162/REST_a_00674. [6]

Chetty, R. et al. (2018), "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility", *National Bureau of Economic Research Working Paper Series*, Vol. No. 25147, http://dx.doi.org/10.3386/w25147. [15]

Citron, D. (2007), "Technological due process", *Washington University Law Review*, Vol. 85, pp. 1249-1313. [23]

Daas, P. and M. Puts (2014), "Social media sentiment and consumer confidence"*, Statistics Paper Series*, No. 5, European Central Bank , Frankfurt. [8]

Dillon Consulting (2018), *Analyzing 211 Rural Unmet Service Needs*, Rural Ontario Institute, http://211ontario.ca/wp-content/uploads/2018/11/ROI-211-Final-Report-Nov-23-2018.pdf. [16]

Epstein, D. and K. Maxwell (2016), *Case study 3: Telamon North Carolina Corporation's collaboration with a county agency*, Child Trends, Bethesda, [20] https://aspe.hhs.gov/system/files/pdf/207836/TelamonCaseStudy.pdf.

Eubanks, V. (2018), *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*, St Martin's Press, New York. [22]

Finkelstein, A. and M. Notowidigdo (2019), "Take-Up and Targeting: Experimental Evidence from SNAP", *The Quarterly Journal of Economics*, Vol. 134/3, pp. 1505-1556, http://dx.doi.org/10.1093/qje/qjz013. [27]

Fiorio, L. et al. (2017), *Using Twitter Data to Estimate the Relationship Between Short-term Mobility and Long-term Migration*, ACM, New York, NY, USA, http://dx.doi.org/10.1145/3091478.3091496. [10]

Global Pulse (2016), *Integrating Big Data into the Monitoring and Evaluation of Development Programmes*, http://Integrating Big Data into the Monitoring and Evaluation of Development Programmes. [30]

Harron, K. et al. (2017), "Challenges in administrative data linkage for research", *Big Data & Society*, doi: 10.1177/2053951717745678, p. 2053951717745678, http://dx.doi.org/10.1177/2053951717745678. [3]

Klein, T. and S. Verhulst (2017), "Access to new data sources for statistics: Business models and incentives for the corporate sector"*, OECD Statistics Working Papers*, No. 2017/6, OECD Publishing, Paris, https://dx.doi.org/10.1787/9a1fa77f-en. [11]

Langedijk, S., I. Vollbracht and P. Paruolo (2019), "The potential of administrative microdata for better policy-making in Europe", in Crato, N. and P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation - How Access to Microdata is Transforming Policy Design*, Springer Open, Cham, https://doi.org/10.1007/978-3-319-78461-8. [29]

Lazer, D. et al. (2014), "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, Vol. 343/6176, p. 1203, http://dx.doi.org/10.1126/science.1248506. [12]

Marchetti, S. et al. (2015), "Small Area Model-Based Estimators Using Big Data Sources", *Journal of Official Statistics*, Vol. 31/2, pp. 263-281, https://doi.org/10.1515/jos-2015-0017. [5]

Meyer, B. and N. Mittag (2019), "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net", *American Economic Journal: Applied Economics*, Vol. 11/2, pp. 176-204, http://dx.doi.org/10.1257/app.20170478. [26]

Ministerio de Desarrollo Social (2018), *Registro Social de Hogares de Chile*, Ministerio de Desarrollo Social, Subsecretaría de Evaluación Social. [19]

Ohlberg, M., S. Ahmed and B. Lang (2017), "Central planning, local experiments - The complex implementation of China's Social Credit System"*, China Monitor*, Mercator Institute for China Studies, Berlin. [13]

O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing Group, New York, NY, USA. [25]

Osoba, O. and W. Welser (2017), *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*, RAND Corporation, Santa Monika, http://www.rand.org/t/RR1744. [24]

Sakshaug, J. et al. (2012), "Linking Survey and Administrative Records: Mechanisms of [1]

Consent", *Sociological Methods & Research*, Vol. 41/4, pp. 535-569, http://dx.doi.org/10.1177/0049124112460381.

Stats NZ (2018), *Algorithm Assessment Report*, https://www.data.govt.nz/assets/Uploads/Algorithm-Assessment-Report-Oct-2018.pdf. [21]

TAKE-Project (n.d.), *Methodology*, https://takeproject.wordpress.com/methodology/ (accessed on 6 August 2019). [28]

Thornby, M. et al. (2018), "Collecting Multiple Data Linkage Consents in a Mixed-mode Survey: Evidence from a large-scale longitudinal study in the UK", *Survey Methods: Insights from the Field.*, https://surveyinsights.org/?p=9734. [2]

Waugh, A. et al. (2018), *Health and Homelessness in Scotland*, The Scottish Government, Edinburgh, https://www.gov.scot/publications/health-homelessness-scotland/. [17]

Yaouancq, F. and M. Duée (2014), *Les sans-domicile en 2012 : une grande diversité de situations*, Institut national de la statistique et des études économiques, Paris, https://www.insee.fr/fr/statistiques/1288519?sommaire=1288529. [14]

Zagheni, E., I. Weber and K. Gummadi (2017), "Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants", *Population and Development Review*, Vol. 43/4, pp. 721-734, http://dx.doi.org/10.1111/padr.12102. [9]