December, 2016

The Potential of Linked Administrative Data for Advancing Homelessness Research and Policy

Dennis P Culhane, *University of Pennsylvania*

The Potential of Linked Administrative Data for Advancing Homelessness Research and Policy

Dennis Culhane

University of Pennsylvania, Philadelphia, USA

› **Abstract**_ *Administrative data enable planners, researchers and programme evaluators to examine service use on a population basis, longitudinally. The potential for linking administrative data across multiple systems further allows for a more comprehensive view of various subpopulations, their patterns of service use and associated costs. Because people who experience homelessness are often users of multiple systems, and are often homeless in part because of ineffective programmes and insufficient aftercare, these data may be crucial to identifying the gaps that need to be filled to prevent and reduce the duration of homelessness spells. These data can also make visible what may otherwise be hidden and understudied aspects of the homelessness problem. The use of administrative data for purposes for which they were not originally intended also raises some ethical issues that must be considered in light of their potential to be misused.*

› **Keywords**_ *Homelessness, administrative data, integrated data systems*

ISSN 2030-2762 / ISSN 2030-3106 online

Introduction

Recent research from Denmark (Benjaminsen and Andrade, 2105), Wales (Welsh Government, 2015), Scotland (Hamlet, 2015) and Australia (Parsell *et al.*, 2016) demonstrates that there is a growing interest in the use of administrative data to study homelessness. Some countries in Europe, including Denmark, Scotland, Wales and France, as well as provinces in Canada (Manitoba and Ontario) and states in Australia (New South Wales and Western Australia) have gone so far as to establish archives of administrative data to support a broad variety of population and social policy research. The use of administrative data for research is not new, but their value in the social sciences is being increasingly appreciated because of their longitudinal nature. Given that they track the services and expenditures of public agencies, they have also drawn interest from governments for their policy relevance. Concerns about privacy protections and data security that may have inhibited data access in the past are being addressed through technological innovations, standardised governance structures and secure work flow processes. As these 'integrated data systems' (or IDS) become more available, access to administrative data could become much more routine, and could enable a new generation of research on populations with complex needs, like those who experience homelessness. In this article, I describe how the environment for administrative data-based research is changing, consider some of the opportunities that this might present for homelessness research, and reflect on some cautions regarding the ethics of this approach.

The Emerging Context for Administrative Data-based Social Science Research

Administrative data are the records collected by organisations to track their routine 'business' activities, including everything from banking transactions through social service contacts to medical treatment records. They can also include registration systems, like those for vital statistics (birth and death records), disease surveillance, and the like. The population and health registries in Scandinavia have been well known and appreciated by researchers for decades. Enthusiasm for 'open data' has put increasing pressure on governments to push vast troves of these administrative data out to the broader public, including data on crime, real estate transactions and environmental quality measures, for example. Some think that access to these 'big data' and the power of new computing approaches, such as machine learning, hold great promise for advancing our understanding of the world.

The legal protections afforded to health, human service, education and earnings records (think patient, client, student and taxpayer information) have meant that these particular administrative data have been largely excluded from these types of public release. Indeed, more often than not, legislative bodies are adding protections and restrictions for access to such data out of a growing concern about the potential breach of confidential information, the vulnerability of these data to identity thieves, and their possible abuse by commercial firms. However, the potential value of these data, particularly through their linkage of cohorts of individuals across policy and service domains, and over potentially long periods of observation, has become increasingly appreciated. The fact that these data already exist, do not rely on self-report, and are population-based has drawn interest from demographers and economists. At the same time, government surveys in many countries are suffering from lagging response rates and high costs, so administrative data are viewed as offering an appealing and more economical alternative to gauge social, health and economic trends. Social innovators also see the potential to use administrative data to track policy experiments at low cost and in real time, improving the speed of the 'knowledge-to-practice' development cycle. Given these potentialities and converging interests, it is not surprising that some governments and researchers have collaborated to figure out how to improve access to these data while strengthening data security and privacy protections.

Integrated Data Systems (IDS) have emerged as a systematic approach to the archiving of sensitive and protected administrative data in some countries, and in several states and localities in the United States (Culhane *et al.*, 2010). The typical data in an IDS can include vital records, immunisation and disease registries, school attendance and achievement, child welfare services, juvenile and adult justice data (arrest, incarceration, probation), housing assistance (including homelessness programme use), income supplements, social services for groups with disabilities, employment and earnings, retirement status and long-term care data. The model IDS creates a secure data infrastructure, enables linkage of records belonging to the same individual over time, and establishes mechanisms for data access. An IDS also usually includes a formal governance structure with representation from the various data owners and custodians. The governing board typically establishes the policies and procedures of the IDS, reviews requests for data access, and reviews study results prior to their dissemination. This 'systematizing' of administrative data storage and access has, in most cases, increased protection of these sensitive data, especially where previous policies were unclear or less fully articulated.

Technology innovations have also improved the security of records and protection against unintended uses (Jones *et al.*, 2014). The encryption of records using parallel linkage keys can mean that personal identifiers do not have to be sent to

the data archive from the source agencies. Computer scientists are also working on methods for creating research datasets from the different source systems 'on the fly' or on a request-by-request basis, meaning that a linked data archive is not even required in order to create a research dataset (removing a single hacking target). Systematic 'perturbation' of the underlying data, including, for example, the spatial shifting of addresses (or longitude and latitude coordinates), can provide further protection against re-identification based on location data or other potentially identifying data, like specific dates for a given treatment and diagnosis. On the researcher's end, remote processing of the data has made it possible to provide secure access to research datasets without ever allowing researchers to view individual records (Jones *et al.*, 2014). Instead, researchers send statistical code to the IDS and have only statistical or aggregate results returned. Multiple forms of user authentication can include certificates and biometrics (fingerprint or eye scan) to ensure that only approved users get access to research datasets and for a prescribed period of time. Results can be reviewed for disclosure risk, and cell suppression or other rules applied before release. Even better, advances in 'differential privacy,' which involves the introduction of intentional error in results but within bounded statistical intervals that won't affect interpretation, provides an automated protection against the re-engineering of data to identify individuals. In effect, the results themselves are simulated.

In sum, the emerging IDS enterprise requires a whole set of legal, ethical, technological and procedural standards to be considered and developed, which might heretofore have been unimagined. Done properly, this should lead to increased data security and privacy protections, while enabling greater access to critically important data that can inform us about social policy, the economy and human development.

The Potential Value of Administrative Data for Homelessness Research

Populations with complex needs, such as people who experience homelessness and people who have multiple disadvantages, might stand to gain the most from these efforts because their well-being is dependent on the successful integration of several agencies and service systems. Indeed, the evident failure in the integration of these services directly contributes to problems like homelessness in many cases. Research using administrative data can make it possible to create useable knowledge for the reform and reorganisation of resources that can help prevent or ameliorate such conditions. Indeed, it has been argued that IDS are most useful for examining 'key transitions' (Fantuzzo *et al.*, 2015) of people across developmental stages of the life course, and of people moving from one system to another, including moving out of an institutional setting – all moments of vulner-

ability that have been linked to the onset and duration of homelessness (Metraux *et al.*, 2010). Below, I consider some of the ways in which homelessness policy and practice could engage in and benefit from this emerging innovation in social science research.

Establishing a data collection standard for homelessness services

In order to participate in a record linkage project, it is first important that the homelessness service system has a data capture system in place for registering or tracking use of homelessness programmes. In the US and Canada, and within the EU – for example, in Ireland, Hungary, the Netherlands and Denmark (Poulin *et al.*, 2008; Busch–Geertsema *et al.*, 2104; Government of Canada, 2016), data standards have been established that require communities or municipalities to collect basic client-level information on the users of emergency shelter services, as well as the dates of service use. These requirements generally apply only to shelter or housing programmes that receive public funding, but participation may be encouraged by all programmes in a community, regardless of their funding sources, to enable a more comprehensive picture of the problem. An obvious limitation is that these systems don't capture people who are not users of homelessness services.

Other countries, for example Ireland and Belgium, require people to register as homeless if they benefit from homeless services or want housing priority, and perhaps to re-certify at various intervals (Pittini *et al.*, 2015). These systems have the benefit of identifying people who might avoid homelessness services but who want access to other benefits associated with their status. A limitation is that these systems may not track service use – specifically, programme entry and exit dates. Thus, there may be only a registration date recorded, limiting longitudinal analysis of the homelessness event(s). In either case, the recording of basic client data, including names, birthdates and national identification numbers, is critical to creating the linkage keys to other datasets.

Establishing information systems of this sort is not a simple process. Many service providers are reticent about recording client-level data for a variety of reasons, including concerns that the data may be used for purposes that are not in the best interests of the clients. In order to address these concerns, a clear set of policies and a governance process have to be established. The policies should articulate the purposes of the data collection, the permissible uses of client information, secure procedures for the sharing of client data for research or evaluation purposes, and clear restrictions on the use of the client data for any purposes not otherwise outlined in the policies. Ideally, clients should be provided with notice of these policies, afforded the opportunity to opt out of data collection, and allowed to review or edit their record.

It is noteworthy that record linkage projects can be undertaken on a more *ad hoc* basis even without such homelessness registration or services tracking systems. A recent example comes from Brisbane, Australia. Parsell and colleagues obtained the written consent of a group of formerly homeless tenants in a supported housing project to link their health (emergency, inpatient, mental health and ambulance), criminal justice (police, prison probation and parole) and homelessness service use. Tenants were assured that their identities would not be disclosed and the information derived would not be used to contact them or affect their access to benefits. Sixty-one percent of the tenants provided consent. The records were linked by the Queensland Department of Health Data Linking Authority, and identifiers were stripped from the final dataset. Researchers were then able to compare aggregate service use and costs pre- and post-housing placement, even absent a homelessness registration system or an IDS. (The authors found that aggregate service use declined from an average of $48,217 (AUS) annually prior to housing placement to $20,788 after housing placement, more than offsetting the costs of the housing programme). So, it is possible to take advantage of administrative record linkages without a homelessness services information system, although it is not as likely to be as efficient or robust for policy analysis purposes as maintaining an on-going homelessness services record system that routinely links to an IDS.

Observational studies

The first benefit of having a homelessness services tracking or registration system is its enabling of basic observational studies of the nature and dynamics of the population. The service data can enable researchers to describe the incidence and prevalence of service use in the local or national population and by discrete subgroups (i.e., by ethnicity, sex, age, household status, disability groups, geography). Epidemiological studies can be undertaken of the relative risk for homelessness programme use by these subpopulations, and of the trends observed over time. Such basic statistics can inform the required scale of potential interventions, and the target subpopulations. They can also be used to assess whether interventions are having an impact on the prevalence or duration of homelessness programme use at community level, and whether other exogenous factors (recessions, broader shifts in social welfare policies) are changing the size and composition of the population over time.

Once the homelessness data are linked with other service system data – for example, through an IDS – researchers can similarly conduct observational studies of the incidence and prevalence of contiguous and/or concurrent service system programme usage. For example, rates of admission to homelessness programmes following discharge from jails or other correctional institutions can be measured, as can admission rates among people discharged from hospitals, emergency

rooms/urgent care programmes, detoxification programmes, or youth transitioning from foster care (see, for example, Metraux *et al.*, 2010). In effect, the homelessness programme data, through its linkage to other service system records, can be used to monitor and evaluate the effectiveness of discharge and aftercare practices of other institutions and service systems.

Correspondingly, other service systems can examine how homelessness impacts them, with their own clients as the reference populations. For example, a community might be able to observe the impact of homelessness on the local jail or on the use of emergency medical transport services, thereby motivating collaboration that might mitigate unnecessary and costly services use while improving the well-being of people who are homeless or at risk of homelessness.

Predictive analytics: identifying and refining target populations

The next phase of research that is commonly supported by administrative data is looking for risk factors and interagency service-use patterns that predict entry to or exit from homelessness. A variety of methods are used to look for subpopulations within the overall population experiencing homelessness that could be targeted for preventive interventions. These can include models designed to look for the latent structure within the population (latent class analysis, factor analysis) or theory-driven approaches that seek to disaggregate the population by predefined variables (cluster analysis). A recent study in Denmark (Benjaminsen and Andrade, 2015), for example, undertook such an analysis, identifying the characteristics associated with patterns of homelessness programme use (see also O'Donoghue-Hynes, 2015, for a similar analysis in Ireland). Recent approaches with 'machine learning' have been growing in popularity; in this approach, statistical programmes look for subgroups based on iterative, best-fitting algorithmic models that are relatively agnostic to theory. Whichever approach is taken, the usual goal of this work is to identify predisposing characteristics or service-use patterns that are associated with the onset and duration of homelessness, with the hope that these can be interrupted and homelessness averted or resolved.

Consider, for example, that homelessness prevention programmes might be challenged by sceptics who could argue that many people who apply for or receive prevention services would not become homeless in the absence of those services and that, therefore, scarce resources are wasted. Indeed, a recent randomised control trial of a homelessness prevention programme in New York City (Rolston *et al.*, 2013) found that 92 percent of the people in the control group did not become homeless despite not receiving prevention services, compared with 96 percent of the intervention group that received prevention services. Interestingly, because the intervention cut homelessness rates in half among those served (from 8 percent among controls to 4 percent among those treated), and because the cost of shelter

use is so high in New York City, the intervention still resulted in modest net savings from the intervention overall. Nevertheless, the study did confirm that the vast majority of people who apply for prevention services (more than 90 percent) are unlikely to become homeless, even without the services. Thus, improved targeting could garner better results and greater efficiency, given the high rate of 'false positive' cases that were served. Shinn and colleagues (2013) used assessment data and administrative record linkages to develop a prevention-screening tool that has done just that, improving the targeting of New York's prevention programme by more than 20 percent. This screener included flags based on previous homelessness programme use, behavioural health services use and child welfare involvement, in addition to self-reported factors, such as current pregnancy. Thus, in this example, predictive analytics were used to reduce the inefficiency inherent in homelessness prevention programming, improving the argument for their importance and social value.

Given the limited resources for supported housing in the US, much effort has also been invested in assessment tools and record linkage efforts to identify those among the people who experience *chronic* homelessness that have the highest rates of services use or risk for premature death. The 'Vulnerability Index' and related tools have used direct interviews with clients, allegedly to identify those most at risk of near-term mortality (Community Solutions, 2016). Other communities have linked homelessness programme records with health and justice administrative records to identify the distribution of acute care services use (hospitals and jails) in the population, and then to set thresholds for priority need populations (i.e., the top decile of users) (Flaming *et al.*, 2009). The ethical considerations of these rationing approaches to the scarcity problem will be discussed in a later section, but it is worth noting that administrative record linkages have been used to support the triaging of housing interventions and the prioritising of target populations (and to justify denying access to some housing interventions for some groups).

Intervention testing

A further, and arguably the most mature, use of linked administrative data is for the testing of interventions to assess their effectiveness. Indeed, the most common use of linked administrative data in homelessness research historically has been to assess changes in the use of services associated with the placement of people experiencing chronic or long-term homelessness in permanent supportive housing. Such pre-post study designs have been quite common in the United States, including many *ad hoc* record linkage projects (done only once to evaluate a single initiative or project). In a survey done in 2008 (Culhane *et al.*, 2008), more than 50 such studies were identified in the US. Similarly, recent studies in Australia (the Brisbane study by Parsell *et al.*, 2016, mentioned previously) and Wales (Welsh

Government, 2015) have looked at pre- and post-intervention service use and estimated cost offsets. Generally speaking, these studies have found that a PSH (Permanent Supported Housing) placement is associated with reduced use of services, although sometimes after an initial increase in service use as part of the process of stabilising a person in their housing. Service reductions (such as reduced hospitalisations or incarcerations) have even been found to offset fully the cost of the housing intervention for some subpopulations, particularly those with high pre-housing costs, such as people with severe mental disabilities, people who are aged, or people who have extensive arrest and jail histories. However, studies generally find that cost offsets for supported housing placements (at least as commonly structured) are more modest for people with less severe disabilities and who are non-elderly. Nevertheless, even finding some reductions in acute care service use has helped to make the moral argument for further investments in housing, as people are better served in housing than being left homeless, and they use more appropriate (and less expensive) support services, regardless of overall cost offsets or realisable savings.

Because most of these pre-post study designs are quasi-experimental, have small sample sizes and very often do not include comparison groups, concerns could be raised that selection bias has skewed results in favour of finding positive cost offsets, as tenants with disproportionately higher need are more likely to be enrolled. The limitations associated with small sample sizes and a lack of comparison groups can be overcome with larger study populations and the use of administrative records to create matched control groups, including through propensity score matching (see Culhane *et al.*, 2002). In such studies, administrative records are used to identify groups with common homelessness service use histories, comparable demographic characteristics, similar behavioural health diagnoses and shared aggregate justice system use (jail stays). These matching efforts can strengthen the argument for the robustness of the outcome, relative to what happens to the comparison group, controlling for regression to the mean or other confounders. And in many cases, this improved and more rigorous quasi-experimental approach may be as good a standard as can be achieved scientifically while also meeting community norms for ethical research, which may preclude randomisation of people experiencing homelessness to interventions.

However, it has also been the case that some large-scale randomised controlled trials (RCTs) have been used to examine the effectiveness of homelessness interventions, such as PSH, targeted at people experiencing long-term or chronic homelessness. Most notably, Canada launched a large, multi-city RCT to look at the impact of PSH on homelessness, including use of services pre- and post-housing placement (Stergiopoulos *et al.*, 2015). The study was able to use tenant interviews to track service use (self-report) – an effort that is now being replicated

with linked administrative data (Hwang, 2016). The study found that service reductions and cost-offsets were significant among the high-need subpopulation, but not significant in the moderate-need subpopulation. This is broadly consistent with the results from the quasi-experimental literature in the US.

Advocates of evidence-based policy-making have also been promoting the use of linked administrative data as a cost-effective and timely way to embed randomised policy experiments in the everyday practice of public administration (Haskins and Baron, 2011; Nussle and Orszag, 2015). These advocates argue that administrative data makes it possible to conduct high-speed/low-cost RCTs as a regular business practice to improve social programmes. They point to private sector businesses, like Google and Bank of America, who conduct thousands of experiments every year by varying their products and tracking the impact in real time with administrative data. Of course, changing social policy is more complex (and sensitive) than changing a website or search algorithm. But the translational message is that the prevailing dominance of carrying out social policy experiments like prescription drug trials – with relatively small samples tracked over three to five years at great expense – is being challenged by a new model of shorter, faster knowledge development cycles made possible by embedded experiments using administrative data systems to measure outcomes.

Some Cautions on the Ethical and Scientific Use of Administrative Data

The era of 'big data' is promising increased knowledge at increased speed, but it is also a Brave New World for the social sciences and social policy, fraught with ethical issues as well as new (or newly significant) scientific considerations. Most of these challenges can be confronted and properly addressed, but they also must be forthrightly and transparently established. The scientific community and the social policy community are often viewed with suspicion by the general public because they are perceived as doing secretive work without proper public scrutiny, and that, consequently, people can be victimised by their elitist machinations. Unfortunately, far too many historical examples support these suspicions. If social science and social policy are to benefit from linked administrative data, with all the perceived hazards associated with privacy compromises and with 'Big Brother' watching, public confidence and trust in the enterprise will have to be assiduously cultivated. That must start with an open consideration and discussion of the ethics of this work.

An ethical IDS effort begins with an explicit and transparent governance process. Most public administrative datasets are the legal responsibility of the administering agency so that, typically, these agencies are legally bound to review and approve any request to access their data (Stiles and Boothroyd, 2015). A governance process, usually spelled out in an interagency Memorandum of Understanding (MOU), or some other document, explicitly outlines the process for obtaining agency consent for a given research project, and the minimum criteria (scientific, policy or practice value) that must be met for a project to be considered. The MOU will also usually describe the data security standards that must be adhered to, and the rules and procedures governing access by external entities like researchers. The governance committee might also involve citizen or nongovernment stakeholder participation (i.e., NGO representatives) to assure that the proposed uses of the data are known beyond the internal workings of government or the academy, and meet community standards for value and integrity. Alternatively, an IDS may have a separate community or stakeholder advisory board, with citizen and other stakeholder representation, which provides periodic feedback to the enterprise.

In addition to a governance policy controlling access to data, a policy should be in place requiring that study *results* be previewed by various stakeholders so that nuances in data and interpretation are discussed. Such a preview and discussion of results contributes to a spirit that the enterprise is jointly motivated by all of the participants in a positive, reform-oriented ethos. If, instead, studies are viewed as padding for academic resumes, or as evaluative 'gotcha' efforts to blame agencies for the shortcomings of programmes, agency personnel will be understandably reluctant to participate. Thus, researchers and nongovernmental stakeholders have to respect that, treated insensitively, evaluative results could be perceived as threatening to the careers of public administrators; thus, a clear set of procedures are needed that provide for adequate feedback and buy-in to the results from public agencies.

In general, then, the basic ethical requirement for the operation of an IDS is a communication strategy that is transparent and that engages government data owners along with other community partners with a vested interest in the programmes under study, as well as general citizenry (including sceptics). In this way, people can participate and inform beneficent policies and uses of the data.

As data security protections become more robust, including through data perturbation and remote statistical analysis – approaches that effectively guard against potential re-identification or re-disclosure of personal information, a further ethical consideration is whether government actually has an *obligation* to contribute data to an IDS or to other important analysis efforts. A government agency may argue that they can't participate in data sharing or data integration

efforts because of re-disclosure risks. But these arguments may also be efforts to shield an agency from evaluative assessments, either about a programme's efficacy or the quality of their data. Again, a shared ethos among participants that efforts are motivated by positive outcomes for citizens, and not for blame, could persuade some reluctant agencies to participate. However, the potential for evasive action does raise the issue of whether or not an *assumption* should be made (as a value or ethic) that publicly-financed programme data must be made available to these efforts because of the potential of these data for contributing to the common good. In other words, interagency data sharing should be the default position, unless a specific statutory authority can be cited to show that it is prohibited (some prevailing law that restricts some data from any sharing, even for evaluation or audit purposes). Such a value can further contribute to the transparency of these efforts, and assure citizens that government cannot hide from uncomfortable truths that may be revealed by their data.

As the use of administrative data for social science and social policy research advances, the ethics of the science of these data must also be considered. First, as with social research in general, research based on administrative data has to meet the human subject protections that generally govern scientific conduct at universities, research institutes and in government through 'institutional review board' or 'ethics committee' approval. In general, the use of 'secondary data' as such represents a much-reduced level of risk to studying participants when compared to primary data collection, with the primary risk being the potential re-disclosure of confidential information to unauthorised parties. Indeed, if appropriate and effective de-identification procedures are employed, such research could be argued as *exempt* from institutional review board approval in most cases.

However, in the event that administrative data are being used to evaluate interventions, then the interventions themselves should be reviewed by institutional review boards for the ethical nature of the interventions, if not for the use of the administrative data to evaluate them. In many countries, interventions that represent variations in public benefit administration, and other *enhancements* to benefits as determined by government, may be exempt from institutional review board approval because the primary goal is improved service delivery and not producing generalisable knowledge. In other words, the goal is to *evaluate* public programmes and their administration, not research to benefit the scientific understanding of a problem. This distinction is important with regard to institutional review board jurisdiction in many communities. However, the potential for government's understanding of an 'enhancement' to be inconsistent with the general public's understanding does suggest that some level of review should be considered in some cases, or even required, depending on the nature of the intervention.

A further ethical consideration regarding the use of administrative data in research is that these data were not collected with the intention of supporting research, and they may not be properly suited to that purpose in some cases. An obvious scientific problem is when the data are incorrect or invalid. Data audit routines should look for incomplete variables and out of range codes. More subtly, agency practice may result in variables being coded in ways that do not correspond to their label or original intent. Variables may be used for 'office' purposes in ways that might not be readily apparent to researchers, which is another reason for making sure that studies are done collaboratively and results previewed with administrative staff to reduce the likelihood of misinterpretation or misuse of data.

Another subtler, but quite serious ethical consideration regards the re-use of data for purposes beyond their original intent. This is especially sensitive when derivative data analytic products are used to determine eligibility for programmes or are used to render decisions about client access or the receipt of services – decisions that could have quite real and serious implications. Consider a couple of examples. In the US, machine-learning algorithms are currently being used to inform judges' decisions regarding the sentencing of persons convicted of various crimes. The algorithms are intended to identify the strongest predictors of recidivism given certain sentences, and to remove the bias of judges (Berk and Hyatt, 2015). Variables for race and ethnicity have been intentionally excluded from these algorithms, but former US Attorney General Eric Holder nevertheless expressed concern about these methods (Barry-Jester *et al.*, 2015). According to Holder's critique, while race may be excluded as an explicit variable, race may nevertheless be embedded in the administrative data because the differential experiences by race of people in the US get effectively encoded in the data through their differential contact with public agencies and programmes. Thus, bias may be 'hidden' in the data, which can reinforce bias in sentencing decisions, calling into question the neutrality of the machine-learning algorithms (Crawford, 2016).

In an example closer to the homelessness situation, the use of administrative data to identify 'high service users' as a way of prioritising candidates for permanent supportive housing is similarly fraught with potential ethical considerations. First, these data were not collected for this purpose, and may well exclude important information which, were it included, might well lead to a person being determined as eligible as opposed to ineligible for certain housing programmes. For example, the fact that some data may be missing from a given IDS or other system (i.e., data from an outlying county where the services were received) may mean that persons are being determined as ineligible or at least not 'priority' when in fact their actual, more inclusive records might indicate otherwise. A second concern relates to whether these 'high service use' thresholds actually make sense in determining need, when in fact they reflect service use and not need. This sort of subtle nuance

is ignored when a simplistic decision rule is derived from crude aggregate utilisation measures. Yet, a third consideration concerns the use of these decision rules to justify leaving some people *unserved*. The fact that people who are homeless and in need of housing would not receive housing due to some decision tool, like the Vulnerability Index, doesn't necessarily indicate that the people may be any less eligible *under the law* to receive that housing. Indeed, it should be no comfort that just because there is a decision rule for rationing limited resources, that a fair result is being meted out. Scarcity may drive rationing, but administrative data shouldn't be used to justify it or even to decide the winners and losers of rationing, without a clearly legislated (i.e., publicly deliberated) intent and authorisation to do so, along with appropriate protections against inadvertent denial of eligibility due to data limitations.

The use of government-held administrative data, where these contain highly private information and are often accessible only to privileged researchers, calls out for a transparent and ethical framework to guide the conduct of this work. While some of the ethical issues can be anticipated and forthrightly addressed in the policies and procedures of an IDS or a given record linkage project, not all of the issues are as obvious. For that reason, continued engagement with members of the general public, with sceptics ('devil's advocates') and with others knowledgeable about the use and misuse of statistics can help to provide checks and balances for the maximum protection of the data *and* for the most beneficent uses.

Conclusions

Homelessness research – and by extension people who experience or are threatened with homelessness – could stand to benefit from improved accessibility of administrative data for social policy research. Administrative data offer a potentially low-cost and continuous source of information regarding the prevalence and duration of homelessness in a community over time. Linked with other records, they can allow assessment of the degree to which discharge practices from other social welfare systems are resulting in increased homelessness. Correspondingly, analysts can assess the disproportionate impact of homelessness on other social welfare systems. Interventions intended to reduce rates of homelessness or to expedite stable exits from homelessness can similarly be tracked with administrative records, and the impact on other health and social welfare systems evaluated. These and related efforts offer a new opportunity for the study of homelessness, and for informing changes in policy that could benefit those who are experiencing homelessness or who are otherwise at risk. This new potential for social science and social policy research is not without its own risks, and it requires an explicit and transparent discussion of the ethical considerations of these linked data systems. A strong policy framework and a clear communication and community engagement strategy can provide some significant protections against the potential misuse of these sensitive data. New technological innovations and secure workflow procedures can also be deployed to provide added protections to confidential data and, ultimately, to leverage those data to expand our knowledge of what works best and for whom.

› References

Barry-Jester A.M., Casselman, B. and Goldstein, D. (2015) *Should Prison Sentences Be Based on Crimes That Haven't Been Committed Yet?* [on-line] Available from: http://fivethirtyeight.com/features/prison-reform-risk-assessment/ [11.07.16].

Benjaminsen, L. and Andrade, S.B. (2015) Testing a Typology of Homelessness across Welfare Regimes: Shelter Use in Denmark and the USA, *Housing Studies* 30(6) pp.858-876.

Berk, R. and Hyatt, J. (2015) Machine Learning Forecasts of Risk to Inform Sentencing Decisions, *Federal Sentencing Reporter* 27(4) pp.222-228.

Busch-Geertsema, V., Benjaminsen, L., Filpovic Hrast, M. and Pleace, N. (2014) *Extent and Profile of Homelessness in European Member States: A Statistical Update – European Observatory on Homelessness Comparative Studies on Homelessness* (Brussels: FEANTSA).

Community Solutions (2016) *The 100,000 Homes Vulnerability Index: Prioritizing Homeless People for Housing by Mortality Risk*. [on-line] Available from: http://100khomes.org/sites/default/files/About%20the%20Vulnerability%20Index.pdf [17.12.2016].

Crawford, K. (2016) Artificial Intelligence's White Guy Problem, *New York Times*, June 25, 2016.

Culhane, D.P., Metraux, S. and Hadley, T. (2002) Public Service Reductions Associated with Placement of Homeless Persons with Severe Mental Illness in Supportive Housing, *Housing Policy Debate* 13(1) pp.107-163.

Culhane, D.P., Parker, W.D., Poppe, B., Gross K. and Sykes, E. (2008) Accountability, Cost-Effectiveness, and Program Performance: Progress Since 1998, in: D. Dennis, G. Locke and J. Khadduri (Eds.) *Proceedings from National Symposium on Homelessness Research* (Washington: Department of Health and Human Services and Department of Housing and Urban Development).

Culhane, D.P., Fantuzzo, J., Rouse, H., Tam, V. and Lukens, J. (2010) *Connecting the Dots: The Promise of Integrated Data Systems for Policy Analysis and Systems Reform* (Philadelphia: Intelligence for Social Policy, University of Pennsylvania).

Fantuzzo, J., Culhane, D.P., Rouse, H. and Henderson, C. (2015) Introduction to the Actionable Intelligence Model, in J. Fantuzzo and D.P. Culhane (Eds.) *Actionable Intelligence for Social Policy*, pp.1-38. (New York: Palgrave Macmillan).

Flaming, D., Matsunaga, M. and Burns, P. (2009) *Where We Sleep: The Costs of Housing and Homelessness in Los Angeles* (Los Angeles: Economic Roundtable).

Government of Canada (2016) *National Homelessness Information System*. [on-line] Available from: www.esdc.gc.ca/eng/communities/homelessness/nhis/index.shtml [17.12.2016].

Hamlet, N. (2015) *Measuring Health and Homelessness in Fife.* [on-line] Available from: www.gov.scot/Resource/0047/00476237.pptx [17.12.2016].

Haskins, R. and Barron, J. (2011) *Building the Connection between Policy and Evidence: The Obama Evidence-based Initiatives* (London: NESTA).

Hwang, S. (2016) *Personal Communication* (On file with author).

Jones, K.R., Ford, D.V., Jones, C., Dsilva, R., Thompson, S., Brooks, C.J., Heaven, M.L., Thayer, D.S., McNerney, C.L. and Lyons, R.A. (2014) A Case Study of the Secure Anonymous Information Linkage (SAIL) Gateway: A Privacy-Protecting Remote Access System for Health-related Research and Evaluation, *Journal of Biomedical Informatics* 50 (August) pp.196-204.

Metraux, S., Byrne, T. and Culhane, D.P. (2010) Institutional Discharges and Subsequent Shelter Use Among Unaccompanied Adults in New York City, *Journal of Community Psychology* 38(1) pp.28-38.

Nussle, J. and Orszag, P. (2015) *Moneyball for Government* (New York: Disruption Books).

O'Donoghue-Hynes, B. (2015) *Patterns of Homeless Emergency Accommodation Use in Dublin: How do we Compare?* Paper presented at European Research Conference, Dublin, 25 September.

Parsell, C., Petersen, M. and Culhane, D.P. (2016) Cost Offsets of Supportive Housing: Evidence for Social Work, *British Journal of Social Work* doi: 10.1093/bjsw/bcw115.

Pittini, A., Ghekiere, L., Dijol, J. and Kiss, I. (2015) *The State of Housing in the EU 2015* (Brussels: Housing Europe).

Poulin, S., Metraux, S. and Culhane, D.P. (2008) The History and Future of Homeless Management Information Systems, in: R. Hartmann McNamara (Ed.) *Homelessness in America* Vol. 3, pp.171-180. (Hartford: Praeger).

Rolston, H., Geyer, J. and Locke, G. (2013) *Final Report: Evaluation of the Homebase Community Prevention Program* (Bethesda, MD: Abt Associates).

Shinn, M., Greer, A.L., Bainbridge, J., Kwon, J. and Zuiderveen, S. (2013) Efficient Targeting of Homelessness Prevention Services for Families, *American Journal of Public Health* 103(S2) pp.S324-S330.

Stergiopoulis, V., Gozdzik, A., Misir, A. *et al*. (2015) Effectiveness of Housing First with Intensive Case Management in an Ethnically Diverse Sample of Homeless Adults with Mental Illness: A Randomized Controlled Trial, *PloS ONE* 10(7): e0130281. Doi: 10.1371/journal.pone.0130281.

Stiles, P.G. and Boothroyd, R.A. (2015) Ethical Use of Administrative Data for Research Purposes, in: J. Fantuzzo and D.P. Culhane (Eds.) *Actionable Intelligence for Social Policy*, pp.125-155. (New York: Palgrave Macmillan).

Welsh Government (2015) *Supporting People Data Linking Feasibility Study: Emerging Findings Report – Research Summary, No. 65/2105* (Cardiff: Welsh Government).