The principles of an ideal homelessness administrative data system: lessons from global practice

Dr. Ian Thomas[1], Administrative Data Research Wales, Cardiff University

Dr. Peter Mackie, School of Planning and Geography, Cardiff University

Abstract

Discussions of homelessness measurement methodologies have largely focused on the generation of primary data, for example point-in-time counts. Though there is long standing tradition in the use of administrative data for measuring homelessness, relatively little examination of administrative data as method exists, i.e. the set of socio-technical practices through which administrative data are generated. This paper undertakes an internationally informed review of 50 administrative data systems in order to deconstruct these systems and stage a methodological discussion. Uniquely, the review included systems from other policy fields outside of homelessness, including health and education, in order to learn from wider data practices. The discussion elaborates on six key design considerations driving administrative data systems, including; function; data architecture; data quality; ethico-legal considerations; privacy preservation; and data access and accessibility. To conclude, we outline what an ideal data system would look like in order to improve the potential use of administrative data to measure homelessness and our response to it, but, more importantly, in mobilising data more effectively in order to facilitate research and operational uses of data. The six design elements can inform future homelessness administrative data systems, whilst also sensitising researchers and users of current administrative data to its (socially) constructed nature.

Keywords: Administrative data; homelessness; system design; international; measurement

[1] Corresponding author: ThomasIR2@cardiff.ac.uk

Introduction

Policy makers, practitioners, and researchers across the globe have been highly critical of the current state of homelessness data, with concerns largely focused on the quality or lack of data to enable consistent and comparative measurement of the issue (Busch-Geertsema 2010; Busch-Geertsema et al. 2016). Whilst the methodological focus of homelessness measurement has been on point-in-time counts, and to a lesser extent the use of capture-recapture (Cowan et al. 1988; Williams 2005), there has been an enduring interest in the use of administrative data for homelessness policy and research (Culhane 2016; Culhane and Metreaux 1997). Administrative data—also known as records or registers—are data routinely generated by organisations and can be considered the 'data exhaust' from operational purposes (Hand 2018). Examples of administrative data include records of stays in shelters, or intake screening when entering a homelessness system.

The general issues with administrative data for research and statistics are widely rehearsed, for example poor data quality and the difficulties of using data generated for other (non-measurement) purposes (Connelly et al 2016; Hand 2018), as are their specific application to homelessness (Culhane, 2016; Edgar et al., 2007; Metraux and Tseng, 2017). In contrast to the growing body of critical literature on administrative data as a data source, the aim of this paper is to discuss administrative data as 'method', i.e. a set of socio-technical practices through which administrative data are generated and deployed. The starting point for this paper is therefore how to design a new homelessness administrative data system, rather than assess the virtues and pitfalls of administrative data for measurement more generally. The paper begins with an overview of the evidence base underpinning our discussion, before moving on to examine several core design considerations of administrative data systems that emerged from our review. The paper concludes by proposing the principles of an 'ideal' homelessness data system.

International systems review

The evidence base for this paper comes from a desk-based review and synthesis of 50 international administrative data systems from 9 countries (Table 1). Relevant systems were initially identified by drawing on the knowledge of homelessness sector stakeholders, with

systems being included for review if they involved the gathering and/or the transmission of personal sensitive data. This initial list was then augmented and validated by identifying peer reviewed journals and 'grey literature' relating to empirical analysis of homelessness administrative data. Published analyses were identified by searching Cardiff University's digital library using a set of key terms related to administrative data[2]. Data sources used in these publications were then identified and the systems that generated them added to the review. Primary literature relating to administrative data systems were located from relevant online sources, e.g. user guides and manuals made available on government or software provider websites. These primary materials were supplemented with secondary accounts of data systems, e.g. in empirical research and statistical publications.

The systems included in the review primarily covered homelessness data. However, data systems from other policy areas were also included in order to learn from wider best practice. For example, health care settings tend to have well developed data systems due to the routine production of administrative data, such as medical notes and medical test results. As this paper aims to discuss an ideal data system, it was appropriate to think outside the current data practices across the homeless sector, which can lead to incomplete pictures of homelessness at local and national levels (Busch-Geertsema, 2010; Busch-Geertsema et al. 2016).

Table 1. Homeless and non-homeless administrative data systems which inform the review, split by country

Country	Administrative data systems
United Kingdom	- Supporting People, Wales - Street Homeless Information Network (SHIN) pilot - SSDA903 collection/Looked after children Census** - Housing Stock Analytical Resource for Wales, UK Secure eResearch Platform** - Mainstay - Greater Manchester Tackling Homelessness Information NetworK (GM-Think/M-Think) - In-Form DataLab - HMRC DataLab** - Ministry of Justice DataLab** - Combined Homelessness and Information Network (CHAIN)

[2] Key terms used in literature search included: administrative data, administrative records, data linkage, linked data, record linkage, linked record.

	- Supporting People Client Record System and Outcomes Framework - Scottish statutory homelessness collections - Homelessness Case Level Information Collection & DELTA - Expanded Troubled Families programme** - Dementias Platform UK Data Portal** - Kent Integrated Dataset** - Connecting Care** - COntinuous REcording of lettings and sales (CORE) ** - NHS Scotland Corporate Data Warehouse & Data Marts** - North West London Whole Systems Integrated Care (WSIC) data warehouse and dashboards** - GP Connect** - Care.data - Secure Anonymised Information Linkage databank**
Ireland	- Pathway Accommodation & Support System
Denmark	- Register of users of section 110 accommodation in Denmark
Poland	- Homelessness and housing exclusion (BIWM) Data Standard
Estonia	- X-tee e-Estonia**
Australia	- Specialist Homelessness Services National Minimum Data Set & Validata - Specialist Homelessness Information Platform - e-Wellbeing platform (Part of the Geelong Project)
New Zealand	- Integrated Data Infrastructure** - Individual Client-Level Data**
United States	- Department of Housing and Urban Development homelessness data collections (National) - New York City Coalition on the Continuum of Care Homeless Management Information System (New York) - Chicago Homeless Management Information System (Chicago) - Online Navigation and Entry System (San Francisco) - Clarity - Nevada Statewide Community and Homeless Management Information System (Nevada) - CARES of NY Regional Homeless Management Information System (New York State) - Ohio Human Services Data Warehouse (Ohio State) - Michigan's Statewide Homeless Assistance Data online Warehouse (SHADoW) - Veterans Health Administration Corporate Data Warehouse** - Virginia Longitudinal Data System** - North Carolina School Works** - Knoxville Homeless Management Information System (KnoxHMIS) - Stella P - Kentucky Statewide Longitudinal Data System - StreetSmart
Canada	- Homeless Individuals and Families Information System (National) - Calgary Homelessness Information Management System (Calgary) - Shelter Management Information System (Toronto)

** Non-homelessness administrative data system

As each system operates within a specific context, whether that be policy, legal, or social, we avoided creating typologies of whole systems, instead choosing to deconstruct systems into a series of six crosscutting areas that emerged as important design considerations, summarised in Table 2. The desk-based analysis revealed possible options relating to each design consideration. The following sections of the paper discuss each of these six core design considerations and drawing on examples from international data systems the paper critically reflects upon options relating to each element.

Table 2: Overview of design considerations and options identified from the review of administrative-data systems

Design consideration	Definition	Approaches adopted (Options)
Function of the system	Proposed use of the data outside its original context of generation	- Measurement of homelessness, and the response to it - Research and analysis -Operational integration of data for decision making
Data architecture model	Pattern of data flows within the data system	- Centralised, bringing together data into a single dataset/system - Federated, where data remain with data owner(s) and are brought together when required - Hybrid models combining elements of centralisations and federation
Data quality	Quality of data is a normative judgement based on intended use, however data should be timely, reliable and valid given their context	- Data standardisation/harmonisation - Active monitoring of data quality - Automated validation
Ethico-legal	The ethical and legal considerations when gathering and storing data	- Consent to share and process personal/sensitive data from the person

		- Using legal means to share/process data, e.g. drawing on specific legislation as enablers
Privacy preservation	Mechanisms of maintaining the privacy of personal data being processed by a data system, thereby meeting certain legal and ethical obligations	- Processing (e.g. aggregation) - De-identification of individual/case level data - Sharing personal information, with higher levels of data security, e.g. Trusted Third Party and split file processes
Data access and accessibility	Accessibility relates to making data interpretable to a wide range of audiences of different 'data literacy' levels, whilst access relates to physically being able to work with the raw data	- Digested information, i.e. portals, dashboards, and open data, meta-data - Raw data, i.e. data downloads - Mediated knowledge, i.e. data labs, automated data generation

Function of the system

Whilst this special issue focuses on measuring homelessness, the review of data systems and the wider literature very clearly highlight how measurement is one of three very broad functional uses of administrative data, the other two being research and operational purposes. Kumar (2015) makes a similar distinction between research and practice orientated uses within the context of Integrated Data Systems. However, from the review, the design and functionality of systems that were designed purely for measurement, as opposed to those that actively used data for research, were markedly different—leading to us separating those two functions. Namely, as will be discussed later, the access and accessibility of data was found to be more limited in purely measurement orientated systems.

Use of data for measurement is largely aligned to homelessness prevalence estimation, service activity, and outcome monitoring, often within the context of performance monitoring to guide service delivery and development at local and national levels. As an example of measurement, the United States Department of Housing and Urban Development (HUD) produce a series of annual statistical outputs on the number and characteristics of homeless people in the United States (Henry et al. 2018). Reports are based

on de-duplicated aggregate counts of homeless people within communities receiving funding from HUD, with the aim of monitoring progress in terms of numbers of people experiencing homelessness. At a local level, the data being collected by communities that feed into this larger national system of measuring the prevalence of homelessness is used to generate outcomes measures that provide an indication of the performance of the community to work as a system of services, e.g. the proportion of people assisted who return to homelessness is measured through re-occurrence at a homelessness service. Missing from the United States' homelessness administrative data systems are measures of activity, i.e. details of actions undertaken by services, although this can be inferred from the type of organisation being funded, e.g. street outreach.

Systems of measurement often—though not necessarily always—result in the generation of standardised and rigorous data: in comparison to purely operational data that tends to be highly unstructured. The higher quality data within systems of measurement enable research and evaluation. The Register of users of section 110 accommodation in Denmarkis an example of data collected for use in measuring activity, specifically placements of people in shelters under Section 110 of the Social Assistant Act, that generates standardised data that has been used for research, specifically through its linkage to other data sources (e.g. Benjaminsen 2016; Benjaminsen and Andrade 2015; Nielson et al. 2011). It should be noted however that the Danish approach to national statistics incorporates data linkage through the widespread use of national person registration numbers, and which greatly facilitates this research use: not all nations are as 'data mature' in their ability to operationalize administrative data, even when collected by governments.

A small subset of administrative data systems were designed specifically for the use of data for research. In the United States, the Virginia Longitudinal Data System (VLDS) and North Carolina School Works (School Works) are both examples of 'statewide longitudinal data systems' intended to enable analysis of linked education data. The VLDS and School Works both bring together education data held by state organisations that cover the breadth of schooling, to enable learner pathways to be explored, with the aim of improving student outcomes through research. Though measurement can help guide allocation of resource and monitor current activity, research has an important role in future facing decision making, for example identifying risk factors and predictors of homelessness and what works in ending

homelessness, both of which can help determine what interventions should be funded based on their efficacy and how they should be targeted.

The final function for administrative data systems we wish to draw attention to goes beyond measurement and research, and entails the direct (re-)integration of data into operational decision-making. However, that organisations use (or should use) their own data for decision-making is self-evident: the operational integration of data we wish to highlight is combining of data from multiple sources to expand institutional knowledge beyond its own boundaries in support of operational decisions. Data integration can occur as part of the measurement and research use of data, as the literature on Integrated Data Systems illustrates (Fantuzzo and Culhane 2015); what marks out operational data integration is that data are tied to/or directly impact 'real' people—rather than the more circuitous route through which policy-making impacts people. For example, Pathway Accommodation and Support System (PASS) in Ireland is a shared real-time platform of homeless presentations and bed spaces across the country and is used as the basis for managing access to emergency accommodation. Similarly, the e-Wellbeing system associated with the school based 'Geelong' intervention in Australia was intended to bring together data about young people at risk of homelessness from different sector actors in order to co-ordinate school support staff and community intervention teams (Mackenzie and Thielking 2013).

Data architecture

The term 'data architecture' is used here to refer to the structure of a data system, specifically flows of data through the system, and can be broadly classified as either centralised or federated architecture—see Table 3 for an overview of the different architectures and their sub-types. In a centralised model, data are periodically reported, or 'pushed', to a central location, where they persist. The creation of a central data repository has the benefit of enabling historical analysis and measurement of homelessness in a timely manner, i.e. without having to engage in lengthy data collection exercises in order to answer each new query. Mechanisms for pushing data are either through the extraction of data from one system and depositing it in another, or several different organisations entering data into the same central repository in a 'live' manner.

An example of reporting data to another organisation is the Specialist Homelessness Services National Minimum Data Set in Australia, which collects information on people referred to or accessing homelessness services and is reported by homelessness services to the Australian Institute of Health and Welfare on a monthly basis. In the United States, the Ohio Human Services Data Warehouse is an example of data being reported into a specifically designed data infrastructure, or warehouse, using (semi-)automated updates. The Combined Homelessness and Information Network (CHAIN) in operation across Greater London is an example of a centralised model where several organisations have access to a shared platform where the 'front end' of the platform is partitioned into areas for each service provider, whilst the back-end links to a person's common record. In addition to improving measurement by enabling de-duplication of people to produce unique counts, shared systems can facilitate the use of data beyond measurement and research, to incorporate data into case management. Most of the systems reviewed adopted a centralised approach, though there were a limited number of systems adopting an alternative architecture: 'federated' data.

A federated data system adopts a 'pull' approach to data flows where organisational data remain distributed and are only brought together or integrated for specific uses. The most common approach to federation amongst the reviewed systems was through a hub and spoke model. Data owners (i.e. homelessness service providers) are the spokes, whilst a central 'hub' organises flows of data through/across the federation, known as the 'data broker'. Upon request, data are automatically extracted from systems by the data broker and combined to form a data set for analysis by the data requester. However, data are for single use only, i.e. for the use by the requester, and as such, data within a federated model is not stored outside of the participating organisations' systems in a permanent repository. The X-tee system in Estonia is an example of a completely automated federated system that enables 'live' querying of other agency databases—and forms the backbone of Estonia's 'e-Government'.

The decision to adopt federated models over centralised one has, in the United States at least, been driven by restrictive state laws against the sharing of personal data. An example of the federated model is the Virginia Longitudinal Data System (VLDS), which enables research access to de-identified school/pupil data without exchanging personal data or processing data outside of its original host organisation, thereby working within the confines

of local legislation. Prior to leaving an education data owner's system within the VLDS federation, data are de-identified thereby rendering them linkable but effectively anonymous. All data being extracted under the same data request undergoes the same de-identification, meaning that the same individual can be linked across different data sources.

Table 3. Summary of data models with examples from the review

Model	Sub types	Examples
Centralised: Data 'pushed' to a single location to form a permanent data pool	***Data set:*** Where data are combined to form a single data set	- H-CLIC, United Kingdom - Specialist Homelessness Services Collection, Australia
	Warehouse: Where data are pooled together in a specifically designed data space	- Kent Integrated Data set, United Kingdom - Michigan's Statewide Homelessness Data online Warehouse, Untied States
	Information system: Where data can be accessed simultaneously by different organisations	- Combined Homelessness Information Network, United Kingdom - Homeless Individuals and Families Information System, Canada
Federated: Data are 'pulled' from organisation databases on demand, and are for single use only	***Live federation:*** Organisations can query one another's databases in-real-time	- X-tee, Estonia
	Data broker: Requests for data are managed by a central data broker who is authorised to extract data	- Virginia Longitudinal Data System, United States - North Carolina School Works, United States

Data quality

By their very definition, administrative data are data that are used beyond their original context, for example records in Ireland where data collected on individuals housed in hostels and other emergency housing provision as part of the PASS system are used to produce regional and national statistics on homelessness. Homelessness administrative data are also often pooled from different service providers/organisations, whether these be different

hostels or emergency accommodation providers (as in the case of PASS in Ireland), different outreach teams (as in the case of CHAIN in Greater London), or different local authorities (as with Scotland's HL1 collection). However, idiosyncrasies in personal and organisational practice can negatively affect data quality. As an example, due to time constraints, frontline staff may not enter all personal data fields when completing intake forms, thereby reducing the ability to de-duplicate people accessing services when attempting to count the number of unique homeless. Inaccuracies in data, or a lack of certain data outright, can lead to policy and decision making that either lacks any evidential basis or is misinformed by apparently reliable evidence; it could therefore be argued that data quality is a precondition of the 'ethical use' of data in decision making (World Health Organisation 2017:30).

Standardisation of what data are collected and by who improves the consistency of data across organisations forming part of an administrative data system. Edgar et al. (2007) propose such a core standard for use across Europe to improve measurement of homelessness across and within nations. Similarly, in the United States, HUD require all funded communities to collect the same core standard as part of their local management information systems. Alternatively, data can be harmonised to make different data providers' data conform to a single data standard—after the fact. A case of the latter style of 'data harmonisation' is the Homeless and Housing Exclusion (BIWM) Data Standard in Poland (Wygnańska 2015). The BIWM was an attempt to create a methodology for enumerating homelessness in Poland by combining data from service providers. In the process of creating the BIWM, differences between pre-existing data collection practices and the intended standard required re-alignment of both practitioners understanding and the final standard, illustrating the difficulties and compromises needed when standardising data, particularly across different organisation types, whilst still maintaining participation in such systems. Measurement and research uses of administrative data are facilitated through standardisation and harmonisation as it increases the coherence and coverage of information when data from disparate sources are pooled, for example leading to 'triangulation' of sources to arrive at more reliable estimates of homelessness—i.e. de-duplicate individuals to generate unique counts—or insight into different aspects of homelessness during a given point in time or over a period of time, i.e. the number of people who have experienced

different forms of homelessness within a year or over their 'homelessness pathways' (Fitzpatrick et al. 2013).

Data standardisation without some maintenance of data standards can lead to a slow decline in data quality over time, as working practices develop that can impact data. For example, in many of the Homeless Management Information Systems (HMIS) covering communities in the United States, the lead organisation whose responsibility it is to maintain the HMIS often provides either reports on data quality, or the ability for data inputting organisations to generate data quality reports themselves. The provision of reports on data quality provides a feedback loop between data input and tangible outputs, thereby increasing the salience of the data entry activities of frontline service staff to those same staff, in addition to highlighting data issues prior to any reporting deadlines, providing time to correct these. Alternatively, organisations can work with data collectors directly to improve quality standards, an example being the Data Quality Campaign (DQC) in the United States, which is a not-for-profit organisation that works with education providers and states in order to improve the evidence base on education. Part of the work of the DQC is to improve data standards, along with use of administrative education data as part of the state funded longitudinal data systems (e.g. Kentucky Statewide Longitudinal Data Systems, Virginia Longitudinal Data System, North Carolina School Works).

Across all the data systems reviewed there were varying levels of automation of data quality monitoring, usually with data being validated when 'in transit'. In Australia, data being submitted to the Australian Institute of Health and Welfare as part of the Specialist Homelessness Services Collection are uploaded via Validata, a secure web-portal that generates reports of errors and data quality to enable data providing organisations to re-submit data after addressing these. Though these validation software can lessen certain administrative tasks around checking the completeness of data, they have less of an impact on the transformation of what some have likened to the transformation of service facing staff from care workers to information processors, with care roles increasingly requiring greater levels of data entry, which can change the nature of their interactions with people seeking assistance (Parton 2008, 2009; De Witte et al. 2016).

Ethical and legal issues

Though the re-use of organisational data is widely espoused internationally, at the extreme end leading to the formation of integrated data systems (e.g. New Zealand's Integrated Data Infrastructure), there are serious ethical and legal dilemmas when using data beyond their original context. Data protection laws, in most instances, determine the legal basis for the initial collection and processing of data for administrative purposes when providing services to homeless people; however, they also determine lawful onward use of administrative data. As an example, the Data Protection Act 2018 and the General Data Protection Regulation in the United Kingdom both stipulate lawful reasons for processing data, which include scientific research or statistical purposes, and the legal obligations to the 'data subject' required to be met in re-use. Mechanisms for addressing legal obligations around data re-use are discussed shortly. Though there has been some attempt to mount an ethical(/moral) argument *for* the re-use of data in order to reduce social harms (Jones et al. 2017), we firstly want to touch upon the more tangible ethical issue of data re-use: namely, the infringement of human rights to privacy and the negative consequences for those already at the margins of society when administrative data systems 'go wrong'.

In an ethnographic study of data analytics, Eubanks (2017) draws on case studies of the negative consequences of technology when applied to decision-making, for example how 'false positives', i.e. errors when integrating data, can result in people being denied or having assistance taken away. Though these errors occur infrequently, when they do occur, they can compound the marginalisation of people already at the margins of society. Systems errors can be exacerbated in cultures of data use where decision-making is deferred to algorithms, i.e. where client facing staff do not want to countermand decisions made by algorithms. Eubanks (2017) advocates that designers should consider how new technologies and data systems impact on people's self-determination and agency and poses a gut check for any new system in considering whether such technologies would be tolerated by the population writ large. Across homeless services it is almost universally espoused that data sharing and integration enables organisations to help people; however, were this type of data integration applied to everyone in society, it would likely be branded 'Orwellian' and dismissed as a breach of privacy rights.

As an example of negative public reaction from integrating data, care.data in England was intended to be a system for extracting and linking General Practitioner data across

England with other health and social care data (Hoeksma 2014). However, the scheme was met with negative responses from the public and health practitioners due in part to the potential for data disclosure and the decision to make the system 'opt-out', i.e. assume consent unless told otherwise. Care.data was eventually abandoned, despite the potential to radically change the evidence base for service provision and policy (Godlee 2016). In the various 'post-mortem' examinations of care.data, it has been highlighted that there needed to be greater transparency around the scheme, particularly how people's data were being used and by whom (van Staa et al. 2016). The need for greater transparency in how algorithms and data systems integrate data and operate can help people who are subjects/objects of these systems to question adverse decisions (Alston 2018; Pleace 2007), whilst drawing on consent mechanisms can address power imbalances between homeless people, and those collecting data about them as part of administrative systems. However, there is a complex interaction between ethical practices, such as consent, and addressing legal issues, as is now discussed.

Though the legal 'gateways' through which administrative data can be reused for research and measured vary internationally, the review highlighted three broad mechanisms that were applicable internationally, namely: (1) consent, (2) through legislation, and (3) obligation. Where use of data directly impacts service users, i.e. by being used in case management, consent to share and link data was obtained from the person being supported. The Online Navigation and Entry System in Chicago, as with other HMIS in the United States, operates as a shared case management system across the community, with data being accessible to other organisations involved in a person's care. Consent is asked for data to be 'visible' to different extents on the ONE System, e.g. sharing of all, some, or no information. However, within the EU data protection context, consent is only valid when freely given, which roughly equates to agreement to use of data without fear of repercussion or coercion. It has therefore been argued by the Information Commissioners Office (2019) in the United Kingdom that public authorities should avoid consent where other means of lawfully sharing data can be used. People using public services may feel coerced to provide data to gain 'better' services, with there existing a clear imbalance of power between individuals and service providers. Though the validity of consent can be argued in certain service contexts, it forms one aspect of ethical practice of engaging people in how their data are being used and

addressing power imbalances in service provision. However, where consent is asked, use of other legal means to continue to use data against a person's wishes undermines the practice and validity of gaining consent.

Aside from consent, the other main gateway to enable use of data is through legislation which was particularly the case where the purpose of data was for measurement and research uses. As an example, the Digital Economy Act (2017) in the United Kingdom is a piece of legislation designed specifically to facilitate the sharing and processing of data between public services for the purposes of service improvement, which includes provision for statistical (measurement) and research uses. However, when drawing on legislation to legitimate the sharing and processing of personal information, there are usually still obligations to the people whose data it is in making the processing fair and transparent, and informing them that their data are being used in certain ways. In the case of the statutory homelessness (H-CLIC) data system in England, 'privacy notices' were issued outlining how individual level data would be used through a layered approach involving posters in public places, information leaflets, and electronically placing the notice on local authority websites.

A final mechanism for data sharing is through obligation, usually to a funder, which was drawn on in only a handful of systems. For example, use of the Calgary Homeless Management Information System (CHMIS) is mandatory for all not-for-profit organisations receiving funding from the Calgary Homeless Foundation, with the CHMIS acting as a shared database and case management solution for homelessness services. However, where funded bodies are obliged to provide data, this does not necessarily obviate the need to abide by data protection legislation and other ethical obligations to the people whose data they are sharing. An example of obligation 'gone wrong' was the decision by the Ministry of Social Justice in New Zealand to contractually require funded third sector organisations to collect and report individual level data (Individual Client-Level Data, or ICLD). The decision to mandate ICLD was met with widespread negative response from several stakeholders, to the extent that a review of the data requirement was undertaken by the Office of the Privacy Commissioner (Edwards 2017). The review highlighted that the purpose for collecting data was unclear, and the requirement to collect individual level data could deter people accessing normal 'low threshold' services anonymously.

Privacy preservation

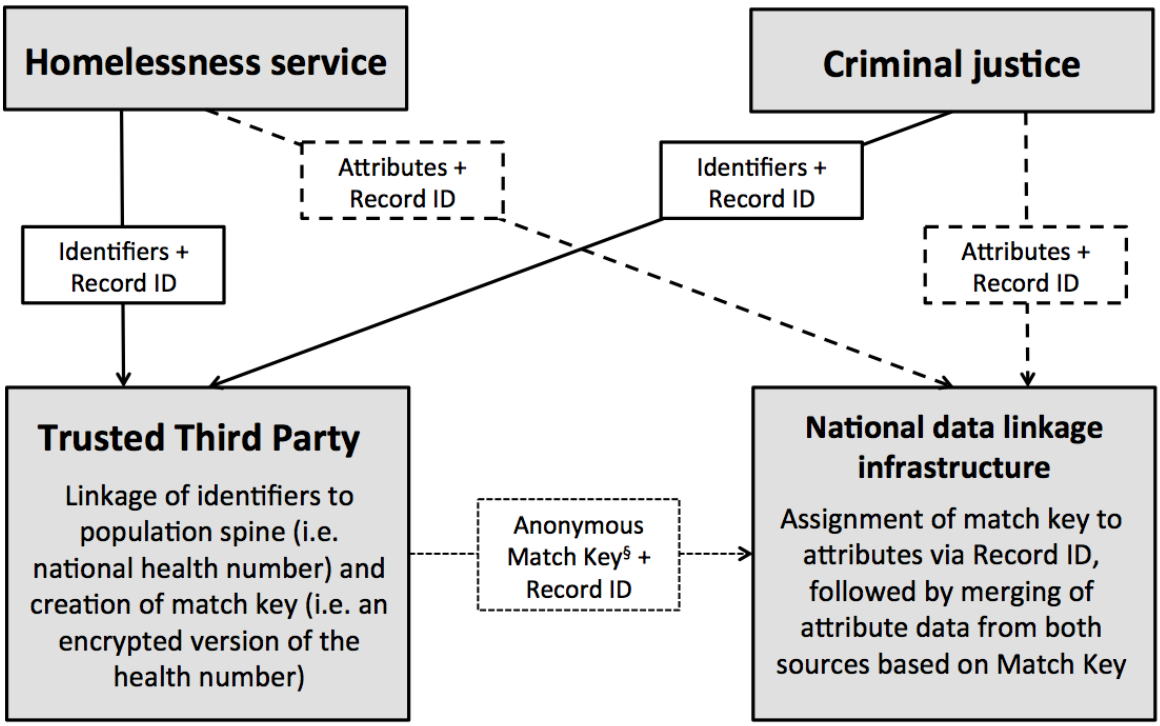
Privacy is an important principle of any data system in increasingly 'surveiled' times (Pounder 2008). Some of the approaches to ensuring the security of data throughout the sharing process are; (1) processing of data, i.e. aggregation, (2) anonymise or de-identify data, or (3) use personal data with added measures to ensure that disclosure risks are minimised, e.g. a 'split file' method of sharing data. Rather than any one of these approaches being 'better', measures taken to ensure the privacy and security of data vary depending on the intended use of data and the local and national legislation around sharing and processing personal data. For example, data as part of operational data-sharing platforms necessitates that people are identifiable given that the purpose of such a platform would presumably be to use the data to make decisions about the case (e.g. HIFIS in Canada and the Calgary HMIS). In cases where data are for statistical or research purposes, the need to measure the same individuals over repeated years, or to link data across sources, was the deciding factor in what method of privacy protection was used. For example, Michigan's Statewide Homelessness Assistance Data online Warehouse (SHADoW) brings together data from numerous homeless Continuum of Care and other public services, in order to provide a research resource. However, without the ability to link between individual records, such a system would not be possible, thereby necessitating the sharing of identifying data.

De-identification of data means that data are effectively anonymised, whilst retaining the ability to link together data relating to the same person; whereas anonymisation would render data un-linkable, which may be appropriate if data are to be used as a standalone resource. When de-identifying data, the same person receives the same unique number throughout the data source, and preferentially across data sources. There were various examples of how the amount of personal data can be minimised or the risks of sharing reduced, including: using a national identification number (e.g. the Register of users of section 110 accommodation in Denmarkuses the person's national Central Personal Register number); creating a unique number based on personal data (e.g. Statistical Linkage Key in the Specialist Homelessness Service Collection (SHSC)/Specialist Homelessness Services National Minimum Data Set in Australia); assigning the person a unique number at random and retaining lookup tables to enable the same person to be assigned the same number if they re-enter a service (e.g. Local Authority Child Identifiers used in the Looked After Children

Census in Wales); and encrypting personal identifiers or already existing unique identifiers (e.g. National Insurance numbers are encrypted before being sent to Scottish Government as part of the statutory homelessness collections HL1 and PREVENT1). De-identification at the source, such as through the creation of unique person numbers, overcomes the risk-averseness of organisations around sharing personal data. However, unless the same method of de-identification is shared with all other data providers participating in a system, the de-identified data will be unlikable.

A 'split file' process can form the basis for national data linkage as it enables the consistent de-identification of data from any data source (e.g Secure Anonymised Information Linkage databank in Wales). Personal identifiers (e.g. name, date of birth, gender, and postcode) are split from data relating to service histories, or 'attributes' (Figure 1). A trusted third party (TTP) receives the personal identifiers and links these to a 'population spine'—a population level set of unique identifiers. National health numbers are widely used as the basis for the population spine in the United Kingdom. Linking to a population spine ensures that the same person, from multiple sources (e.g. homelessness and police data in Figure 1), is assigned the same unique identifier number, and can therefore be linked across all data sources. Unfortunately, this method of de-identifying data does not overcome the initial hurdle of organisations being risk averse, as the personal data still need to be shared with the TTP.

Figure 1: Illustration of hypothetical 'split file' process used to combine data from homelessness services and criminal justice system



Source: Adapted from Harron (2016)

Data accessibility and access

For data to have value, the information generated from it needs to be accessible to a wide range of audiences with varying levels of data literacy, whilst there ideally also needs to be a means via which data can be accessed so that stakeholders can meet their own information/knowledge needs. Drawing on the 'public health' literature, data made accessible to local stakeholders can enhance accountability and action and achieve greater impact than data and analysis that are reported at only national levels (World Health Organisation 2017:32). This emphasises the importance of data and knowledge flowing across boundaries, rather than being siloed by governments and states, or only being the preserve of data analysts and academic knowledge brokers. Three very broad mechanisms were identified for achieving data accessibility and access: (1) digesting the data to generate information (e.g. portals, dashboards, and meta-data), (2) making the raw data available (e.g. for download), and (3) mediating the access of data (e.g. through data labs). However, one commonality was that the use of visualisation significantly improved the accessibility of data,

with this holding true across data at different units of analysis, i.e. individual level up to aggregate data.

When directly linked to local administrative data systems, well designed dashboards can provide easily interpretable 'live' insight into homelessness in an area, which can enable communities to track the impacts of their activities on, for example, returns to homelessness e.g. Los Angeles Homeless Services Authority. Operating at a more granular level were the North West London Whole Systems Integrated Care dashboards that visualise information about individuals in order to reconstruct their interactions with health and social care services; visualisation can then be used by people involved in a persons' care to make future decisions. Increasing the accessibility of data by presenting pre-analysed 'bits' of information can shorten the feedback loop between data generation and action, with this being the driving force behind many campaigns to end homelessness, such as Community Solutions and their Built for Zero campaigns in the United States (Community Solutions 2018).

Access to 'granular' data at the individual or household level is understandably more constrained due to legal and ethical issues around privacy, with access to only de-identified or completely anonymised data available in most cases. Anonymised or de-identified data were found to be accessible either via provisioning of extracts directly to the requester (e.g. Virginia Longitudinal Data System where data from across the federation are compiled and made available for download), or, more often, within a secure data environment (e.g. New Zealand Integrated Data Infrastructure, or Dementias Platform UK Data Portal). Secure data environments can be physically secure spaces, such as the HMRC Data Lab in the United Kingdom, where tax records can be accessed using computer terminals in a 'secure room' based in the HMRC offices in London (Almunia et al., 2019). Alternatively, a secure environment could be a virtual workspace within which all research is conducted and can be accessed remotely via any Internet enabled computer, such as the Secure Anonymised Information Linkage databank 'Gateway' in Wales (Jones et al. 2014). Physical settings are often limited in number, meaning that researchers can sometimes be required to travel to a setting to access data, which may be impractical for some researcher teams; remotely accessible secure environments therefore reduce these access barriers.

Often neglected as a means of making 'granular' raw data accessible, i.e. interpretable, is meta-data, which are data about data and outline the variables contained in

a data set and the values the variables take. Meta-data can be consulted prior to or during research in order to determine the suitability of a data source for a research project, i.e. that it contains the variables needed. However, despite the importance of meta-data in making sense of data, there is a widely acknowledged lack of meta-data that can be a barrier to access and use of administrative data (Jones et al. 2019; van Panhuis 2014), significantly frustrating any use of the data in a timely manner as users of data must spend time understanding the data prior to actual analysis. In addition to making research and analysis a difficult task, a lack of meta-data can also complicate sharing of data between organisations—which speaks back to the need for data standards (themselves a form of meta-data), as a way of achieving a shared 'data language'.

To close, several data systems incorporated knowledge mediators through 'data labs' and software, whose role was to analyse data on behalf of service providers, who may not have the capacity or capability to conduct primary research and evaluation with individual level data. The Ministry of Justice DataLab (MOJ DataLab) in the United Kingdom offers third sector organisations the opportunity to explore the possible associations between their services and their users' recidivism—measured as reconviction rates (Lyon et al., 2015). Organisations submit personal identifiers (i.e. name, date of birth, address) of the individuals taking part in their programme/services to DataLab, who then link this data to the criminal justice data MOJ hold—primarily prison data, but also police data—and generate a comparison group of similar characteristics. A standard report is then generated that compares recidivism rates between people receiving the programme/service and the comparison group. In a more automated approach to mediated knowledge generation, Stella P in the United States is a piece of software developed for use by Continuum of Care that enables them to upload data from their HMIS systems and for that data to be visualised, thereby enabling them to assess the performance of their homeless service system.

Summary: Building an ideal homelessness administrative data system

In relation to each of the six design elements we conclude that in order to improve the use of administrative data to end homelessness, whether through its measurement or through research or practical decision-making, systems should:

- Strive to accommodate measurement, research, and operational purposes in order to maximise the use of data in ending homelessness. At a minimum, a homelessness administrative data system should be flexible enough to evolve over time to meet different functional uses and data requirements.
- Adopt a centralised data architecture model in order to provide a permanent data pool, more likely to be characterised by quality, consistent data, that persists historically, thereby enabling longitudinal measurement and research.
- Use a multi-faceted approach to data quality, combining standardisation, monitoring & automated validation, as this is likely to lead to improvements in quality and consistency, particularly in situations where multiple organisations of differing service provision are providing data. Maintenance of data quality should also be considered integral to any administrative data system, given that poor quality data can impact decision-making—and therefore have serious ethical consequences for homeless people.
- A nuanced combination of consent, legislation, and obligation is likely to be necessary to navigate the ethical and legal collection and processing of individual level personal data of homeless people. Though administrative data are important as a means of measuring homelessness, it is crucial that the requirement to collect data should not deter people from accessing services, nor should it compound the already existing power imbalances between people seeking assistance and homeless services by placing demands upon them.
- Privacy is an important principle of any data system in increasingly 'surveiled' times. However, in many cases in order to engage in accurate measurement of homelessness that eliminates double counting and potentially over-inflates estimates, services and researchers require person level data. Where this is the case, the default should be to de-identify data or share personal data with added measures to ensure that disclosure risks are minimised.
- For data to have value, the information generated from administrative data systems need to be accessible and understood by a wide range of audiences with varying levels of data literacy. To achieve this goal we argue for a combination of approaches, including: the development of data portals and dashboards; making the raw data available for own analyses; and mediating the access of data through data labs that can foster a culture of research and evaluation.

Conclusions

This paper has synthesised international examples of administrative data systems in order to stage a discussion of the 'methodology' of administrative data, which we pose as a series of six areas to consider when designing a new administrative data system. This discussion is important given the increasing international use of administrative data for the purposes of measuring and researching homelessness. In addition to aiding the design of new homelessness data systems, the design elements outlined in this paper provide a way of sensitising researchers and policy makers to the socially constructed nature of the administrative data used to measure homelessness. Decisions made under each of the six design areas outlined create a reality of homelessness-in-data.

Researchers drawing on administrative data from pre-existing administrative data sources should always be wary of the socially constructed origins of these data in organisational practices, which they themselves come to reflect, such that administrative data are not an 'objective' view of the world but embedded in particular institutionalised ways of knowing (Gomm 2004). Decision-making by service facing staff in the homelessness sector will vary within and across organisations due to individual interpretations of policy and practice guidance, whilst their data recording practices will be dependent on workloads (De Witte et al. 2016). At best, administrative data therefore offer one viewpoint of homelessness through the lens of institutions.

Finally, we must reiterate a fundamental limitation of administrative data are their generation from interactions between organisations and people, i.e. where people enter or are entered (potentially involuntarily) into 'systems'. Whilst many people will seek assistance in their homelessness journey, there are others, particularly those with no recourse to public funds, who may not enter any system. From an analytical perspective we could console ourselves that 'at some point' populations appearing in different administrative data systems will likely overlap and are therefore not truly missing; this however is of little practical use if the intention is prevention of harm. The point at which a person enters (homelessness) administrative data systems is arguably too late from a prevention perspective. Therefore, whilst this paper seeks to advance the method of administrative data, we recognise that they need to be part of a wider 'data landscape' on homelessness, one supplemented by other

methods that may be better suited to providing insight on particular populations not 'visible' to institutions.

References

Almunia, M., Harju, J., Kotakorpi, K., Tukiainen, J. and Verho, J. (2019) Expanding access to administrative data: the case of tax authorities in Finland and the UK, International Tax and Public Finance 26 pp.661-676

Benjaminsen, L. (2016) Homelessness in a Scandinavian welfare state: The risk of shelter use in the Danish adult population. Urban Studies 53(10) pp.2041-2063

Benjaminsen, L. & 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

Busch-Geertsema, V. (2010) Defining and measuring homelessness, In: E. O'Sullivan, V. Busch-Geertsema, D. Quilgars, & N. Pleace eds. Homelessness Research in Europe, pp. 19–39. Brussels:FEANTSA

Busch-Geertsema, V., Culhane, D., and Fitzpatrick, S. (2016) Developing a global framework for conceptualising and measuring homelessness, Habitat International 55 pp.124-132

Community Solutions (2018) Getting to the proof points: Key learning from the first three years of the Built for Zero initiative. [Online] Available at: https://community.solutions/new-report-reflections-from-the-first-three-years-of-built-for-zero/ [Accessed: 24 November 2019]

Culhane, D. (2016) The Potential of Linked Administrative Data for Advancing Homelessness Research and Policy. European Journal of Homelessness 10(3) pp.109-126

Culhane, D. and Metraux, S. (1997) Where to from here? A policy research agenda based on the analysis of administrative data. In: eds. D. P.Culhane and S. P. Hornburg, Understanding Homelessness: New Policy and research perspective (Washington, DC: Fannie Mae Foundation)

De Witte, J., Declercq, A. and Hermans, K. (2016) Street-level strategies of child welfare social workers in Flanders: The use of electronic client records in practice. British Journal of Social Work 46 pp.1249-1265

Edgar, W., Harrison, M., Watson, P. and Busch-Geertsema, V. (2007) Measurement of Homelessness at European Union Level (Brussels: European Commission)

Eubanks, V. (2017) Automating Inequality: How high-tech tools profile, police, and punish the poor (New York: St Martin's Press)

Fantuzzo J., Culhane D.P. eds (2015) Actionable Intelligence (New York: Palgrave Macmillan)

Fitzpatrick, S., Bramley, G. and Johnsen, S. (2013) Pathways into Multiple Exclusion Homelessness in Seven UK Cities. Urban Studies 50(1) pp.148-168

Godlee, F. (2016) What can we salvage from care.data? British Medical Journal 354(i3907) pp.1

Gomm, R. (2004) Social Research Methodology: A critical Introduction (Basingstoke, UK; New York: Palgrave Macmillan)

Hand, D.J. (2018) Statistical challenges of administrative and transaction data. Journal of the Royal Statistical Society 181(3) pp.555-605

Harron, K. (2016) Introduction to Data Linkage. (Essex, England: Administrative Data Research Network)

Hoeksma, J. (2014) The NHS's care.data scheme: what are the risks to privacy? British Medical Journal 348(g1547) pp.1-3

Jones, K.H., 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 pp.196-204

Jones, K.H., Laurie, G., Stevens, L., Dobbs, C., Ford, D.V. and Lea, N. (2017) The other side of the coin: Harm due to the non-use of health-related data. International Journal of Medical Informatics. 97 pp.43-53

Jones, K.H., Tingay, K.S., Jackson, P. and Dibben, C. (2019). The Good, the Bad, the Clunky: Improving the Use of Administrative Data for Research. International Journal of Population Data Science 4(1) pp.1-11

Kumar P. (2015) An Overview of Architectures and Techniques for Integrated Data Systems Implementation. In: Fantuzzo J., Culhane D.P. (eds) Actionable Intelligence (New York: Palgrave Macmillan) pp.105-123

Lyon, F., Gyateng, T., Pritchard, D., Vaze, P., Vickers, I. and Webb, N. (2015) Opening access to administrative data for evaluating public services: The case of the Justice Data Lab. Evaluation 21(2) pp.232-247

Mackenzie, D. and Thielking, M. (2013) The Geelong Project: A community of schools and youth services model for early intervention. Report. (Swinburne: Swinburne Institute for Social Research, Swinburne University)

Metraux, S. and Tseng, Y. (2017) Using Administrative Data for Research on Homelessness: Applying a US Framework to Australia. The Australian Economic Review 50(2) pp.205-213

Nielson, S.F., Hjorthøj, C.R., Erlangsen, A. and Nordentoft, M. (2011) Psychiatric disorders and mortality among people in homeless shelters in Denmark: a nationwide register-based cohort study. The Lancet 377(9784) pp.2205-2214

Pleace, N. (2007) Workless people and surveillant mashups: Social policy and data sharing in the UK. Information, Communication & Society 10(6) pp.943-960

Pounder, C.N.M. (2008) Nine principles for assessing whether privacy is protected in a surveillance society. Identity in the information society 1(1) pp.1-22

van Panhuis, W.G., Paul, P., Emerson, C., Grefenstette, J, Wilder, R., Herbst, A.J., Heymann, D. and Burke, D.S. (2014) A systematic review of barriers to data sharing in public health. BMC Public Health 14(1144) pp.1-9

van Staa, T., Goldacre, B., Buchan, I., and Smeeth, L. (2016) Big health data: the need to earn public trust. British Medical Journal 354(i3636) pp.1-3

World Health Organization (2017) WHO Guidelines on Ethical Issues in Public Health Surveillance (Geneva: World Health Organization)

Wygnaska, J. (2015) Measuring Homelessness and Housing Exclusion in Poland: The BIWM Data Collection Standard. European Journal of Homelessness 9(1) pp.123-144