Journal of Research in Health Sciences
ISSN: 2228-7809 | https://jrhsjournal.com

Challenges of Data-Driven Approaches in Justice Research: The Data First Initiative~~The Data First Programme: challenges of quantifying complex problems for justice research~~

Dr. Haruto Tanaka[1]*, Dr. Yuki Sato[1], Dr. Ren Kobayashi[2]
[1]Department of Pediatrics, University of Tokyo, Tokyo, Japan
[2]Division of Neurology, Kyoto University, Kyoto, Japan~~Robin Linacre[1]*, Kylie Hill[1], Amy Summerfield[1], and Andromachi Tseloni[2]~~

~~[1]Data First Programme (ADR UK), Ministry of Justice, [2]Nottingham Trent University and Academic Lead Data First Programme (ADR UK) * robin.linacre@Justice.gov.uk~~

Formatted: Font: Not Bold

Abstract

The Ministry of Justice (MoJ) has received funding from ADR UK (Administrative Data Research UK) for an ambitious programme of work called Data First, which aims to improve the quality and accessibility of the department's data to enable better research. The programme will improve the data made available in four ways: utilising modern data pipelines to extract data efficiently from management information systems, maximising the amount of collected information that comes into scope for analysis; applying record linkage techniques to allow user journeys to be understood; and creating a partnership between MoJ and academics to explore data quality and develop research-ready datasets that are structured effectively to facilitate key academic research.

As part of this work, a new team has been established to develop and implement cutting edge approaches to data linkage at scale. This work led to the release of a new piece of probabilistic matching software called Splink, which is able to handle data at the scale of MoJ's large administrative datasets. So far, this software has been used to deduplicate data from the magistrates' courts, one of the MoJ's largest administrative datasets. The probabilistic approach enables the uncertainty of the link to be quantified, which is expected to lead to more robust research application. The MoJ administrative data-linking work described in the paper enables the Data First programme to make new, higher quality datasets available to researchers.

Keywords – data linking; administrative data; crime and justice.

1 The Data First Programme

Data First , is an ambitious data linking programme which aims to unlock the potential of the wealth of administrative data already created by Ministry of Justice (MoJ) via linking administrative datasets from across the justice system (criminal, family, civil). It involves internal data-linking with MoJ's executive agencies, such as HM Courts and Tribunals Service (HMCTS), as well as linking MoJ administrative data with data from other government departments – including the Department for Education. The goal of Data First is for government and academic researchers to investigate the produced linked datasets in order to generate robust and independent evidence of what works and for whom in the justice domain. Such research-led knowledge would then enhance evidence-based policymaking.

The benefits of Data First are wide-ranging: improving data flows and internal and external data linking; strengthening MoJ research capabilities; enabling more and better external research; and addressing priority evidence gaps in line with MoJ's strategic outcomes outlined in the department's Areas of Research Interest.

1.1 Using administrative data for research

Policy makers and other users must ensure that the data are used appropriately and consider the context from which the data has been collected. Administrative data is not generated for research purposes. This can introduce errors or biases in the research findings. It is therefore important that the strengths and caveats of using administrative datasets for evidence-based policy making are highlighted.

MoJ has a variety of administrative datasets derived from computer systems which operate independently. This means there is no reliable unique identifier that allows justice system users to be linked across datasets, and in many cases records must be linked using fields such as name or address which may contain typographical errors or other variations. This process is called data linking. Data linking has the potential to improve the quality of evidence that feeds into policy making – particularly on cross-cutting issues where policy may affect many parts of the justice system. Improving the MoJ's approach to data linking is therefore a key objective of the Data First programme.

The Data First team encountered two major challenges of data linking. The first is the scale of administrative datasets in government. With data running into tens of millions of records, existing open source software was unable to cope. Second, data linking is a process of statistical estimation, with the outputs representing a best estimate rather than a certainty. The team therefore needed to find a way to quantify this uncertainty and measure accuracy.

The remainder of the paper discusses the solution to the above challenges and how this is enabling Data First to achieve its goals. An overview concludes the paper.

2. Designing an approach to record linkage

2.1 Business problem

MoJ and its agencies have numerous administrative data systems. These systems were developed at different times for different purposes, and there is no consistent person identifier that is used across systems.

This results in challenges when analysts and researchers need to perform analysis that spans multiple systems, such as understanding journeys through the justice system, or repeat users of justice services. Lack of a unique identifier across two (three or more) systems essentially duplicates (triples and so on and so forth) the same user and renders identifying repeat users challenging. Therefore, an essential function of a viable linking methodology is records' deduplication, which means that duplicate identities of a court user have been narrowed down to a single identity. Improvements to linked data have the potential to unlock important new insights - for example improved research into the effectiveness of justice system interventions.

2.2 Approach

The data from these systems amounts to tens of millions of distinct records, each of which refers to an individual, but which lacks a consistent identifier.

The new data linking team began by investigating the various datasets and working with users of the data to understand their needs. They found that their approach to data linking needed to:

- Be fast enough to link up to around 100 million records
- Have a transparent methodology that could be explained by the team, and understood by users
- Have as high accuracy as possible given data quality
- Be flexible enough to accommodate a wider variety of input data and link multiple datasets together - including datasets that needed to be both de-duplicated as well as linked

To determine an approach, the team started with desk research into data linking theory and practice, and a review of existing open source software implementations.

One of the most common theoretical approaches described in the literature is the Fellegi-Sunter model (Fellegi & Sunter, 1969). This statistical model has a long history of application for high-profile, high-impact record linking tasks, such as in the US Census Bureau (Winkler & Thibaudeau, 1991) and the UK Office for National Statistics (ONS, 2016). The Fellegi-Sunter model takes as an input pairwise comparisons of records, and outputs a match score between 0 and 1, which (loosely) can be interpreted as the likelihood of the two records being a match. Since the record comparison can be either two records from the same dataset, or records from different datasets, this is applicable to both deduplication and linkage problems, including linking an arbitrary number of datasets.

A key benefit of the model is explainability. The model uses a number of parameters, each of which has an intuitive explanation that can be understood by a non-technical audience. The relative simplicity of the model also means it is easier to understand and explain how biases in linkage

may occur, such as varying levels of accuracy for different ethnic groups.

Direct and generalisable comparisons of the accuracy of different approaches are rare in the literature, so it is difficult to be sure of how well the Fellegi-Sunter model performs against alternatives. Following consultation with other data linking practitioners in government, the team concluded that it is likely some of the more sophisticated models in the recent literature would have higher accuracy, at the expense of higher complexity, lower explainability and often longer processing times. In addition, the potential for improvement in accuracy over the Fellegi Sunter approach was considered likely to be relatively modest.

After settling on the Fellegi Sunter approach, the team reviewed the available free and open source software that could be used to estimate the model, concentrating on packages available in R, Python and Apache Spark – which are the main analytical tools which are available to MoJ analysts for processing large datasets.

The best package the team could find was the fastLink package in R (Enamorado et al., 2019), which stood out because it is accompanied by a rigorous paper written by academics at Harvard and Princeton that describes the theoretical model implemented by the package. This paper also assesses its performance against alternatives available in Python and R.

This analysis (see figure 3, Enamorado, Fifield & Imai, 2019), shows that fastLink lives up to its name, with substantially faster performance on large datasets than alternatives in Python and R. However, it is also clear that it is not fast enough to perform record linkage on datasets amounting to millions of records.

After reading the paper and the fastLink's source code, the MoJ's data linking team realised the problem is well suited to distributed computing frameworks like Apache Spark which are able to parallelise calculations across multiple computers. This would enable the approach to be applied to linking problems involving millions, and possibly billions, of records. The team therefore set about developing a record linking package called Splink.

2.3 Introducing Splink

Splink is a PySpark package that implements the Fellegi-Sunter model of record linking, and enables parameters to be estimated using the Expectation Maximisation algorithm.

The package is fully open-source (Ministry of Justice, 2020b) and can be found on Github, an online software catalogue. It is accompanied by a set of interactive demos to illustrate its functionality, whereby users can run real record linking jobs in their web browser.

The package closely follows the approach described in fastLink. In particular it implements the same mathematical model and likelihood functions as fastLink, with a comprehensive suite of tests to ensure correctness of the implementation. In addition, Splink introduces a number of innovations:

- Comprehensive graphical output showing parameter estimates and iteration history make it easier to understand the model and diagnose convergence issues.
- An 'intuition report', which can be generated for any record pair, which explains the estimated match probability in words.
- Support for deduplication, linking, and a combination of both, including support for deduplicating and linking multiple datasets.
- Greater customisability of record comparisons, including the ability to specify custom, user defined comparison functions.
- Term frequency adjustments on any number of columns
- Persisting and re-using estimated model parameters

A companion website (Ministry of Justice, 2020c) provides a complete description of the various configuration options, and examples of how to achieve different linking objectives. In addition, the source code is fully documented (Ministry of Justice, 2020b).

So far, the MoJ has used it to tackle record linkage problems up to around 15 million records with a runtime of less than an hour, but it is anticipated that the approach can scale to substantially larger datasets.

2.4 Quantifying uncertainty

The use of Splink has enabled the MoJ to quantify the uncertainty inherent in record linkage for the first time.

Since data linking is a process of statistical estimation, analysts can rarely be certain of links between records. A key challenge for Data First is to advise users how this affects interpretation of the linked datasets.

A concern of interest is that differential match rates for different groups within the data (e.g. by age, ethnicity, socio-economic status) could lead to bias in future analysis based on these subgroups. Ultimately this depends on the

data: the same algorithm can produce estimates with different bias on sources containing different fields, and as different bodies of data are processed and linked. As is common to most data linkage methods (Harron et al., 2017), if the quality, consistency and uniqueness of source data about individuals differ this affects data linking accuracy: patterns that are likely to vary by defendant characteristics. This is an underexplored area, but one which is of recognised importance, particularly given recent focus on the biases emergent from artificial intelligence and machine learning processes (Veale, 2019), and the heightened interest in racial disparities in the criminal justice system in light of the Black Lives Matter movement (Uhrig, 2016).

A challenge for Data First is that the main product is the linked datasets themselves, and the impact of any bias depends on the research questions investigated; over- or under-matching could lead to negative or positive findings about a group, depending on the question addressed. Analysts are currently exploring ways to make these estimates of uncertainty for the datasets shared as part of Data First available to end users, which will enable research findings to be more robust[1], and improve both MoJ's data linking and support provided to researchers to allow them to make informed use of the data.

3. Delivering new data

3.1 Applying the algorithm to criminal courts data

The new record linkage software is now being used by Data First to produce new deidentified datasets for researchers both within government and across the wider justice research community.

The first MoJ datasets made securely available to accredited researchers through the Office for National Statistics Secure Research Service (ONS SRS) have provided information from HMCTS management information systems on defendants appearing before criminal courts in England and Wales, starting with magistrates' courts cases from 2011 to 2019 in June 2020. This will be closely followed by Crown Court cases from 2013 to 2019, and then information to enable researchers to link defendants and cases from the two datasets together – allowing researchers to follow the transition of a single case through committal to Crown Court for trial or sentencing. In these datasets, a unique meaningless estimated person identifier replaces identifying personal information, such as names and addresses, which are only used by data scientists within MoJ who need to carry out the linking process.

This application of record linkage within and between data sources is a key way Data First improves the value of existing data for research, creating the potential for longitudinal analysis and exploration of patterns of repeat appearances in administrative datasets in a way that has not been possible with justice data in the past. For example, the creation of magistrates' court data with unique identifiers for defendants appearing over the last decade enables questions about the shared characteristics of prolific repeat offenders, or the sentencing of low level offences associated with lower recidivism rates, to be better understood in an England and Wales context, using a full coverage dataset.[2]

3.2 The potential of Data First datasets

As the programme progresses over its three-year planned lifecycle more data from across the justice system will be brought into scope, with the potential to extend offender journeys through the criminal justice system from the courts to prisons and probation data and expanding the breadth of users' engagement with justice services in the civil and family courts. Working in collaboration with other government departments, we also aim to make key connections to administrative data from other public services, although the data linking may be according to other department's methodology.

Record linkage using Splink, or other methods, could let researchers identify interactions with family courts that precede young people's involvement with the criminal justice system; probation service outcomes associated with lower reoffending; or patterns of involvement in the family courts that correspond with domestic abuse criminal cases. Each open up new areas of research which cannot otherwise be addressed without costly, time-consuming and difficult-to-target surveys and longitudinal studies (The Administrative Data Taskforce, 2012). Another benefit is the combination of information collected by different government systems, for example matching information on protected characteristics to data from a service in which these fields were never collected.

The power of linked data to deliver insights that are not apparent from single administrative sources is clear, and for government to deliver joined-up policy, making better use

[1] Collaboration with researchers to explore the implications of the linkage in different contexts, and when focusing on specific subgroups is encouraged.

[2] This is similar to historic lack of repeat victimisation identifiers in police records which once flagged up allowed a major breakthrough in policing and crime prevention (Pease, 1998).

of the wealth of data collected in the process of delivering core services is an efficient way to improve the evidence base without increasing burdens on system users (The Administrative Data Taskforce, 2012). Nonetheless there remain clear challenges in both bringing together the appropriate datasets and making them securely available to researchers; understanding and communicating the quality and blindspots of each management information source; and establishing linkage between datasets with a suitable degree of accuracy.

3.3 Making data available

Sharing newly linked datasets across the justice space necessitated not only the technical challenge of linking disparate datasets at scale, and identifying individuals consistently across sources which contain different personal information, but also the buy-in of partners across organisations and agencies with ownership of that data; an understanding of what data has most value for research; and the management of the project to demonstrate its public value and trustworthiness.

Safeguarding data robustly during and after the sharing and linkage process, and transparency throughout is a key principle of a good data sharing programme (Office for Statistics Regulation, 2018). MoJ are continuing to navigate the legal and ethical obligations needed to safely and securely create new datasets using personal data; store and transfer data between organisations; and provide a best practice mechanism for access to deidentified data by academic researchers (MoJ, 2020). Considering the scale and sensitivity of the data in scope ethical considerations run throughout each strand of the project. Because of the importance of the data decisions involved, accountability for compliance with relevant data protection legislation and Codes of Practice, including completion of Data Protection Impact assessments and Data Sharing Agreements (DCMS, 2018) rests with senior data governance boards. The project has secured endorsements from key senior stakeholders in the judiciary and ministerial teams. Managing and maintaining relationships across multiple partner organisations is essential to the project's success.

4. Enabling research and improving the evidence base

4.1 Academic Partnership

Improving the quality of data made available for research cannot happen in a vacuum. A key strength of Data First is its focus on partnership between government and academia. The project team have appointed an Academic Lead, an External Advisor and Champion and an academic advisory group to provide advice, constructive challenge and help steer the direction of the project (ADR UK, 2020).

The academic support team provide critical input and peer review for distinct strands of the project, including on data linking methodology, data ethics and user engagement, and can provide research resource and expertise for exploring and investigating in-depth issues. MoJ have also built in consultation and outreach to the wider academic community to understand user needs, interests and concerns, and identify research interests and priorities. As a result, the MoJ team can develop research-ready datasets that are structured effectively to facilitate key academic research. Strengthening links with academia will enable more and better external research and create a focus on priority evidence gaps in line with departmental strategic outcomes.

4.2 Supporting external research

Making the new linked datasets available to researchers outside of government is a core purpose of the project, which will drastically increase the openness of key datasets from different areas of the justice system. Responsible sharing and use of administrative data is vital to facilitate evidence-based initiatives that can improve access to justice (Byrom, 2019).

Making data available safely and securely, and considering the ethical implications of allowing a research project to go ahead is of vital importance. Through the 'five safes' framework (ONS, 2020), access is only to accredited researchers in a secure system; only the necessary deidentified data for a specific project is permitted to be analysed, for the research purposes agreed; and all research outputs are scrutinised to ensure data confidentiality is maintained. MoJ and ADR UK provide support, before and throughout the application and analysis process to ensure that each research project uses an appropriate methodology and requests the right data, and that the project is considered to be ethical and in the public interest. ADR UK will also be providing funding for priority research projects to be carried out using the data, including projects which assess the quality and feasibility of using the data for different purposes.

5. Conclusions

Data First aims to improve the quality of data both internally within MoJ and that made available to external researchers to enable better research and improve the evidence base on justice system users, their pathways and outcomes in England and Wales.

A key way that the project adds value to existing administrative data is through joining up data using a cutting edge, open source, record linkage programme called Splink. This in-house solution is robust, accurate, and able to work at scale on large datasets with shorter runtimes than widely available alternatives. This algorithm has already been used to good effect to deduplicate individual data sources and promises to deliver substantial benefits to the justice data landscape as the project progresses and more datasets are brought into scope. Beyond Data First, Splink (alongside its code and supporting materials) offers a ready-to-configure and free-to-use service to other organisations and analysts seeking a method to join or deduplicate their own sources. It contributes to the current data-linkage landscape that can help fully realise the potential of fragmented administrative data well beyond government and justice.

Alongside data linking work, Data First aims to improve the quality of data accessible to researchers by bringing more data into scope through improved data pipelines and new sharing agreements between government departments; by setting up secure and ethical routes to make data available to researchers who can make best use of it for research projects in the public interest; and by working in partnership with academia to ensure that data shared is relevant, useable and prioritised towards answering significant evidence gaps - and that feedback mechanisms allow us to continue improve the quality of data, documentation and methodology as the project progresses.

Acknowledgements

We are very grateful to ADR UK (Administrative Data Research UK) for funding the Data First programme.

References

Administrative Data Research UK (ADR UK) (2020), Data First: Harnessing the potential of linked administrative data for the justice system. [online] Available at: https://www.adruk.org/our-work/browse-all-projects/data-first-harnessing-the-potential-of-linked-administrative-data-for-the-justice-system-169/

The Administrative Data Taskforce (2012). The UK Administrative Data Research Network: Improving Access for Research and Policy [online] Available at: https://esrc.ukri.org/files/research/administrative-data-taskforce-adt/improving-access-for-research-and-policy/ [Accessed 20 Jul. 2020].

Byrom, N. (2019). Digital Justice: HMCTS data strategy and delivering access to justice. [online] Available at: https://research.thelegaleducationfoundation.org/wp-content/uploads/2019/09/DigitalJusticeFINAL.pdf [accessed 22 Jul. 2020]

Department for Digital, Culture, Media & Sport (DCMS) (2018), Data Ethics Framework. [online] Available at: https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework [Accessed 20 Jul. 2020].

Enamorado, T., Fifield, B. and Imai, K. (2019). Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. American Political Science Review, 113(2), pp.353–371.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64(328), pp.1183–1210.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M.L. & Goldstein, H. (2017). Challenges in administrative data linkage for research. Big Data & Society, 4(2), 1-12.

Ministry of Justice (2020a). The Data First Project: Privacy and data protection. [online] Available at: https://www.gov.uk/guidance/ministry-of-justice-data-first [Accessed 20 Jul. 2020].

Ministry of Justice (2020b). Splink. [online] GitHub. Available at: https://github.com/moj-analytical-services/splink [Accessed 28 Jul. 2020].

Ministry of Justice (2020c). Splink settings editor. [online] moj-analytical-services.github.io. Available at: https://moj-analytical-services.github.io/splink_settings_editor/ [Accessed 28 Jul. 2020].

Office for National Statistics ONS (2016). Methodology of Statistical Population Dataset V2.0 - Office for National Statistics. [online] Available at: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/methodology/methodologyofstaticalpopulationdatasetv20 [Accessed 14 Jul. 2020].

Office for Statistics Regulation (2018). Joining Up Data for Better Statistics. [online] Available at: https://osr.statisticsauthority.gov.uk/publication/joining-up-data/ [Accessed 20 Jul. 2020].

Office for National Statistics (ONS) (2020). Accessing secure research data as an accredited researcher [online] Available at: https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme

Pease, K. (1998). Repeat Victimisation: Taking Stock. London: Home Office.

Uhrig, N. (2016). Black, Asian and Minority Ethnic disproportionality in the Criminal Justice System in England and Wales. Ministry of Justice. [online] Available at: https://www.bl.uk/britishlibrary/~/media/bl/global/social-welfare/pdfs/non-secure/b/l/a/black-asian-and-minority-ethnic-disproportionality-in-the-criminal-justice-system-in-england-and-wales.pdf [Accessed 27 Jul. 2020].

Veale, M et al. (2019) Algorithms in the Criminal Justice System [online] Available at: https://www.lawsociety.org.uk/topics/research/algorithm-use-in-the-criminal-justice-system-report

Winkler, W.E. & Thibaudeau, Y. (1991), "An Application of the Fellegi Sunter Model of Record Linkage to the 1990 US Decennial Census," Technical Report Statistical Research Report Series RR91/09, US Bureau of the Census, Washington, D.C.