Evaluating the Risk of the Troubled Families Programme Public Data Release

201374125

Introduction

The Troubled Families Programme aims to provide a network of support for families suffering from multiple and complex problems, including parents who do not view work as an achievable goal, with children who have a serious risk of disadvantage (Department for Communities and Local Government, 2017).

This report concerns the publication of data associated with this programme for research purposes. Particularly, the goal with this data release is to evaluate the performance of this programme, and identify areas for improvement. The data is described in the *National Evaluation of the Troubled Families Programme 2015-2020* Report by the Department for Communities and Local Government. **Table 1** gives an overview of the family data provided as part of this dataset, used in an evaluation of the programme. Each dataset concerns the same set of families, and smaller numbers are due to non matching to additional data. National Impact Study (NIS) data may be either unmatched, or matched to administrative datasets, including the Police National Computer (PNC), the National Pupil Database (NPD), and the Work and Pensions Longitudinal Study (WPLS), see **Table 2**.

Table 1: Overview of the data used in the National Evaluation of the Troubled Families Programme report.

	No. of individuals	**No. of families**
National Impact Study (matched and unmatched)	253,230	63,671
National Impact Study (matched only)	187,097	61,664
Family Progress Data	230,858	58,566

Table 2: Overview of the linked data combined with the Troubled Families Programme data.

	No. of individuals	**% of individuals matched**
National Pupil Database (NPD)	92,759	84.3
Work and Pensions Longitudinal Study (WPLS)	74,635	76.1
Police National Computer (PNC)	29,824	18.6

Assessment of the Data

This report will first identify particular data anonymisation issues through reference to the Anonymisation Decision-making Framework (ADF) (Elliot *et al.*, 2016). As the data is still under full confidentiality, this report assesses only the structure of the data as outlined in the *National Evaluation of the Troubled Families Programme 2015-2020* Report (Department for Communities and Local Government, 2017). Particularly, this report considers the first key components from the ADF:

1. Describe your data situation
2. Understand your legal responsibilities
3. Know your data
4. Understand the use case
5. Meet your ethical obligations

Data Situation

Figure 1 gives an overview of the data flow as proposed by the research initiative. At present, the data is only accessed in house. Family and individual level demographic data is provided by local authorities, and matched to administrative datasets held by Government departments, including Police National Computer (PNC), held by the Ministry of Justice, The National Pupil Database (NPD) held by the Department for Education and the Work and Pensions Longitudinal Study (WPLS), held by the Department for Work and Pensions. In this case, these institutions are considered to be the data controllers, see **Figure 1**. The programme data at present is provided in house to the researchers within the *Ministry of Housing, Communities & Local Government*, it is proposed to be made available to researchers to carry out deeper analysis on the programmes effectiveness. The data is at present entirely constrained within government organisations, and is constructed by the processor using no open access data.

Legal Responsibilities

The data is required to comply with the General Data Protection Regulations (GDPR) (European Union, 2016), including the focus on personal data, defined as:

> *"an identified natural person is one who can be identified, directly or indirectly."*
>
> — **GDPR 2016/679 Article 4(1)**

and anonymous data:

> *"anonymous information, namely information which does not relate to an identified or identifiable natural person"*
>
> — **Recital 26**.

Particularly there are key revisions to personal data laws brought forward through the GDPR:

1. Revised Definition of personal data, including indirect and direct identification
2. Consideration of pseudonymisation

Of particular interest is the concept of pseudonymisation by which data which may appear anonymised may still be identifiable as personal data:

> *"Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on on identifiable natural person"*
>
> — **Recital 26**

The data in question does not contain directly identifiable information such as names or addresses, but considering the large amount of indirect personal information provided, it is possible that the data itself is pseudonymised.

Understanding the Data

In order to ensure the data provided gives both a suitable level of depth for the purpose it is intended, and exclude unnecessary information which may be used for reversing anonymisation, this section explores how

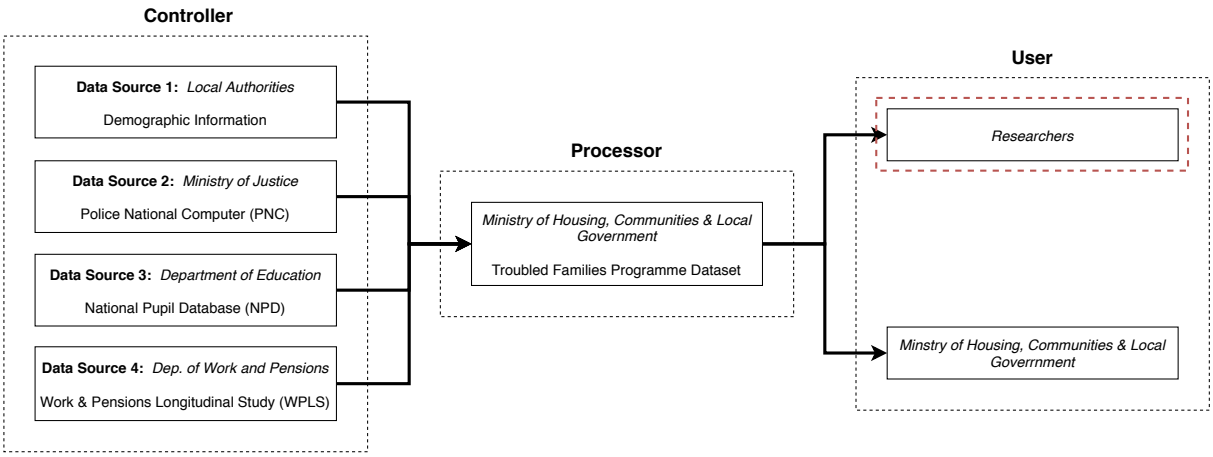

Figure 1: Proposed data flow indicating the extension with red dotted line.

the data may be manipulated to conform with the legal and ethical responsibilities in relation to personal data.

Given this data contains the data linkage of a variety of indirect personal information from different closed sources, there is a particular concern that reversing the anonymisation of this data would be trivial (Harron *et al.*, 2015). The purpose of this data release is purely to allow researchers to perform more in depth analysis on the effectiveness of the Troubled Families Programme. As such, it is likely that the data may be simplified, to ensure there is less risk to anonymisation.

This report considers the first two aspects of the four stages of ensuring anonymity described in Elliot *et al.* (2016), both of which do not require access to the data:

- Data Minimisation from the data situation audit
- Scenario Analysis
- ~~Data Analytical Approaches~~
- ~~Intruder/Penetration testing~~

Data Minimisation

The full dataset in question concerns over 60,000 families, across the United Kingdom. It is likely that a smaller sample of the data would still provide the same level of analytical detail. However, as the data provides a very small subset of the total population, it is unlikely that any personal information could be directly inferred due to the sampling size.

It is unlikely that all variables need to be kept, major identifying variables such as the age of parents and children may be considered for removal. Any unnecessary demographic information, if included, such as religious views, country of birth should also be removed. Additionally, school level information may not be entirely necessary, for example, the specific school children are attending may not be necessary for the analysis. Table 3 gives an overview of some select variables, it should be noted that while some variables may be considered medium or low risk of not conforming with anonymity, when taken together they may be considered a higher risk. For example, if a person is identifiable, their work status may be considered high risk personal information, and when taken together, religion, work status, and location may give identifiable information for persons in unique circumstances.

Table 3: Overview of variables expected to be found in this data. Type of variable indicates whether it is considered to be directly identifiable, risk indicates the risk of the variable not conforming with ethical or legal anonymisation, and importance indicates the level of importance in relation to the goal of the data release.

Variable	**Type of Variable**	**Risk**	**Importance**
Name	Direct	High	Low
Age	Indirect	High	Medium
Religion	Indirect	Medium	Low
School Attended	Indirect	Medium	Low
Location	Indirect	High	Medium
Police History	Indirect	High	Medium
Work Status	Indirect	Low	Medium

The most straightforward method for ensuring anonymisation would be to aggregate the data. Aggregation may be to either family level (rather than individual), or to a standardised geographic area, for example, to census tracts. In this sense, anonymisation is achieved through generalisation, as described in k-anonymisation (Sweeney, 2002). Additional methods may consider the banding of age groups rather than complete removal. For example using 20 - 30, instead of a specific age. Given are examples of three levels of aggregation:

(1) 40 year old male, history of domestic violence, lives in L5, Liverpool, never worked, married to 40 year old wife, with 3 children aged 7, 15, 17, attending local school.

(2) Family of 5, three children, domestic violence, Liverpool, no income.

(3) Liverpool L5: Average number of children = 3, Proportion of families with domestic violence = 30%. 20% Unemployed.

Based on the original data concerned in this programme, (1) gives an example of the sort of individual level information that may be derived. While (1) contains no direct personal information, Elliot *et al.* (2018) note that a person attempting to reverse anonymisation may have access to an unpredictable amount of additional information. For example, the 40 year old male may have a social media accounts containing all the personal information not included here, which may be obtained through searches based on this published data. Through this therefore, a person may discover a known, identifiable person, and determine that they have a history of domestic violence.

Similarly with (2), as outlined in the GDPR, it is likely that pseudonymisation may be a particular concern in some instances of this data. For families in unique circumstances, e.g. with a large number of children of specific ages in a particular area, it may be trivial to reverse the anonymisation. This relates to the GDPR *means reasonably likely to be used* test, given the time, cost, and technical capabilities of such an exercise would be low both when the data is combined processor stage, and when publicised for researchers.

It should be noted that if any directly identifiable features are present in the data, these should be excluded through complete suppression. Irrelevant information such as religion may also be excluded through this method (Sweeney, 2002).

Scenario Analysis

This data contains closed information regarding past criminal offences of many individuals involved, as such reverse anonymisation may be attempted in order to de-anonymise this sensitive information. In terms of the resources required for this, at its current state the data provides a large amount of information that would simply need to be linked with online social networks to provide personal information. As this data is proposed to be open access for research it would not be a problem for anyone to carry out such an attack.

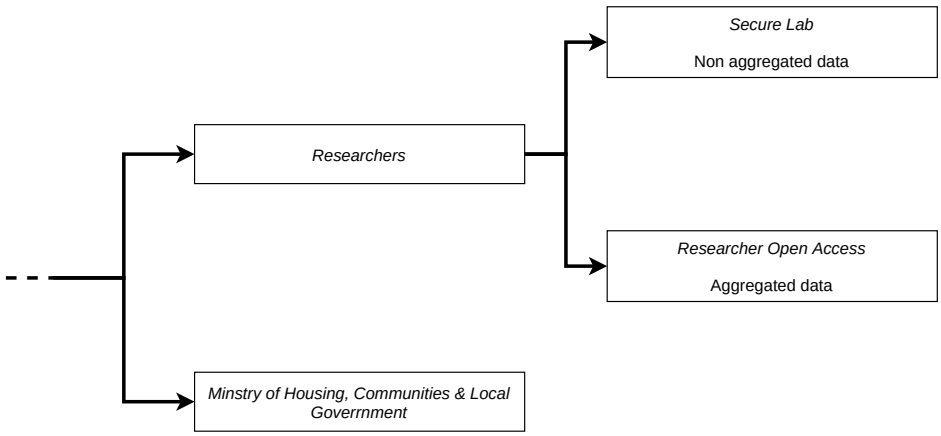

Figure 2: Proposed extension to the researcher accessed data

It is worth considering that data at an individual level is rarely open access, and demographic information such as the census are required to always be aggregated when publicised.

Introducing Controls

Finally, another solution to ensuring anonymisation is to restrict open access in some way. Restricting access follows environmental controls put forward by Duncan *et al.* (2011) by restricting 'who' may access the data. Elliot *et al.* (2016) note that this is essentially about agent control, and by restricting access to 10 people, rather than allowing 10,000 to openly access the data, the risk level is decreased. Consideration must be made as to whom may access the data therefore if given restricted access. In this case, it seems likely that access may be designated to trusted research institutions, those of which have a history of ethical use of restricted data.

This particularly relates to the overall goal of this data release, in that the goal is purely to provide improved analysis of the effectiveness of the Troubled Families Programme. Governance control in this regard may restrict the analysis permitted, and in particular would ensure that when publishing any results of analysis, the data involved is not indirectly accessible for anyone who was not permitted access in the first place (Elliot *et al.*, 2016).

In order to select a balance between full, open access to the data, and a high quality dataset for in-depth analysis, the data may be distributed using a combination of the control measures and data minimisation discussed. **Figure 2** demonstrates an approach to solving the issues by providing aggregated open access data for research, while also permitting certain institutions full unaggregated access, ensuring that some high quality data is made available.

Conclusion and Recommendations

The data in question is comprised of a number of closed access data sources, each of which contain large amounts of high risk personal data in their raw form. Even with the exclusion of any direct personal data, the aggregation of these datasets have the potential to provide indirect methods for reversing anonymisation through relatively simple means, either solely or with the assistance outside, uncontrolled data sources.

The recommendation of this report is to ensure that the data primarily is stripped back to its core variables of interest in the ultimate goal of assessing the effectiveness of the troubled families programme. This includes the removal of any high risk variable, e.g. exact ages, unnecessary police report information, schooling information, or work history. Additionally, the data should not be made available at an individual level for

anyone, instead aggregated to a family level, as should be expected for an assessment of a family targeted programme. This data should not be made available openly, instead contained closed to select institutions through a lab. While open access to the other data may be provided through an aggregation to LSOA census units.

References

Department for Communities and Local Government (**2017**) *National evaluation of the Troubled Families Programme 2015-2020: Family outcomes -national and local datasets: Part 1.*

Duncan, G.T., Elliot, M. & Salazar-González, J.-J. (**2011**) *Statistical Confidentiality: Principles and Practice.* New York, NY: Springer New York.

Elliot, M., Mackey, E., Editor, C.T. & O'Hara, K. (**2016**) *The Anonymisation Decision Making Framework-Mark Elliot,* p. 171.

Elliot, M., O'Hara, K., Raab, C., O'Keefe, C.M., Mackey, E., Dibben, C., Gowans, H., Purdam, K. & McCullagh, K. (**2018**) Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review.* 34 (2). pp. 204–221.

European Union (**2016**) *REGULATION (EU) 2016/ 679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 27 April 2016 - on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation).*

Harron, K., Goldstein, H. & Dibben, C. (**2015**) *Methodological developments in data linkage.* John Wiley & Sons.

Sweeney, L. (**2002**) K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.* 10 (05). pp. 557–570.