PROTOCOL VERSION NUMBER AND DATE

Version 1.0 Date: 1 April 2021

NIHR 133680

Research Protocol

Title: Unlocking data to inform public health policy and practice

Funding:

This study is funded by the National Institute for Health Research (NIHR) [**PHR Project:NIHR133680**]. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Background: The 1998 Crime and Disorder Act requires police, local government, and the NHS to collaborate on joint crime reduction strategies that includes data sharing to inform targeted responses. Violence has been further prioritised by HM Government in the Violent Crime Strategy [1] and the UK government has allocated funds for the formation of Violence Reduction Units [Violence Prevention Partnerships, VPU in South Wales], in 18 police forces with the explicit purpose of promoting a Whole System Multi-Agency (WSMA) approach [2]. Information sharing is vital as it has been shown that often one agency does not have the whole picture, for example 75% of those attending Emergency Departments with assault related injury are not found in police records due to fear of repercussion [3] This means there is a real move to improve collaborations with the police to promote data sharing [4] and recent research has been undertaken to examine the best way of using police data in order to benefit from the 'big data' investments and more efficient decision making [5]. This work focuses on unlocking data with a focus on police data for the prevention of domestic abuse. In this work domestic abuse is defined as any incident of controlling, coercive or threatening behaviour, violence or abuse between those aged 16 or over who are or have been intimate partners or family members, regardless of their gender or sexuality. The abuse can encompass but is not limited to the following types of abuse: psychological, physical, sexual, financial, emotional. [6]. According to Home Office Statistics 1 in 4 women has experienced domestic abuse and 1 in 5 sexual assault during their lifetime in the UK [7]. The Crime Survey of England and Wales reported that 20% of women (3.4 million female victims) have experienced some type of sexual assault since the age of 16 [8]. The abuse often gets worse during pregnancy and this puts both the mother and the unborn child at danger and it increases the risk of potential negative foetal and mother outcome including miscarriage, infection, premature birth, and injury or death to the baby [9]. Pregnancy can trigger or increase the risk of domestic abuse [9,10] and for this reason the Antenatal Routine Enquiry into Domestic Abuse was introduced across Wales in 2005 and it is now a requirement that all women are asked about domestic abuse at a safe opportunity during their antenatal and health visiting appointments. The threat of abuse has been exacerbated substantially in the current COVID-19 pandemic situation [11,12]. World Health Organisation has called on multiple organisations including Governments and policy makers, healthcare facilities and providers, humanitarian response organizations, community members and the women who are experiencing abuse to respond to this crisis during the pandemic.

This work is initiated by Public Health Wales and aims to identify vulnerable women at the stage of pregnancy or within the first year of life in order to offer early interventions and preventative support and help and prevent adverse child events, need for being taken into care and improve outcomes for families.

Pilot work and experience: Secure Anonymised Information Linkage (SAIL) databank is a world-leading secured data linkage, management and access system which holds vast amount of routinely collected population level data from wide range of data providers including National Health Service (NHS), Welsh Government, Office of National Statistics (ONS) and other social care organisations[13] . SAIL facilitates the secured linkage mechanisms across all the datasets held in the databank by assigning a unique double encrypted linkage key (Anonymous Linking Field (ALF))to individual records [14]. SAIL has ISO27001 accreditation [15] and already hold datasets on, looked after children, children in receipt of care data, maternal and child indicators, National Child Health Database, education, higher education and training, family justice data, health (GP/hospital/A&E/intensive care) and others. Bringing data from the Police together with these datasets would enable the police and public health services to offer more evidence based [16] prevention and identify vulnerable families much earlier. However, the barriers to bringing police data with other data sources are greater than in other fields. The main barriers include:

- Barrier 1 - Data type: the data collected by the police is often detailed narrative/qualitative data. This is very rich data but not in a form that can be easily anonymised or coded to extract information.
- Barrier 2 - Multiple software systems: each police force has purchased different software systems for collecting the same type of data (e.g. Public Protection Notification Report (PPN) data is collected using a different system in each of the areas of Wales (North Wales, Powys have different systems, with South Wales & Gwent having NICHE). In addition, there are different datasets within each area using different software systems ware not linked together. The datasets are normally developed locally with discussion with force analysts, and there is no nationally agreed data set to report back on to the Home Office. Even within each area the datasets are not linkable, for example, linking datasets holding data on PPN, Police National Computer, MARAC (Multi-agency Risk Assessment Conference) within the local area is not possible. In addition, some systems are proprietary, such as CCTV data use can require permission from the data software provider to access and decode the data. This can make data sharing even more complicated.
- Barrier 3 - Security of data: The police culture is that they hold highly sensitive data which needs to be held highly secure. There is not a general culture of sharing databases with others outside the force [17]. There is sharing of individual cases under MASH (Multiagency Safe Guarding Hub) but there is little experience or knowledge of what benefits there would be if datasets were linked with other agency datasets. Therefore, the background is there is a high risk in sharing sensitive data but little tangible evidence of the benefits.
- Barrier 4 - Correct identification of individuals: The individuals who are held on police datasets are often highly mobile and move areas regularly, have different addresses and don’t always register with a GP when they arrive in a new areas (which is the method used in Wales to identify residence). There is a police personal identification number but it is not clear how easy this would be to link this with NHS number or pupil ID number in Education datasets or identification in social services. For this reason, methods of confirming the correct individuals are linked across datasets is needed.

Research Questions: This work aims to pilot real work solutions and learn by doing. The work is to maximise the use of data to facilitate cross sector working including public health, police, education, social services, courts, charities and local authority.

Solution to barrier 1 – Can text mining methods be used to extract and code information such as if a person was referred for Police Watch or not. Therefore, enabling rich text information to give coded flagged information which can be integrated with other data sources.

Solution to barrier 2 – What needs to be in place to harmonise different software systems collecting data for the same purpose? Is it feasible for different forces to use the same single system (e.g. NICHE for PPN's enabling comparisons across forces and with other agencies) or is it possible to select data for harmonisation in a core minimum dataset? For example, in Cwm Taff the Multi agency Safeguarding Hub (MASH) use a software product called MHUB which has been developed for each agency to add their updates to each case and all have access. However, this solution still requires each agency to add updates to the system and does not combine existing datasets. This means there is duplication of effort and only specific cases are shared.

Solution to barrier 3 – Can an exemplar case study illustrate the benefits of sharing data among agencies in a safe secure environment with staff who are data security trained. This work will link PPN data for households containing a pregnant women, with data including the maternal and child indictor data, health data (GP/hospital - for mental health/medication), substance abuse dataset, education. This work will show how linked data can help profile vulnerable families better for police and public health support and intervention.

Solution to barrier 4 -Scoping of best methods of tracking people through the electronic system and understanding the errors in linkage or level of non-linkage.

Research Plan and Methods:

Work package 1: Text mining. Supervision Professor Spasic.

Police called to domestic abuse incidence have the ability to issue on the spot protection notices. These notices provide temporary protection to the person who is subject to the abuse. The notification can prevent a person from contacting the aggrieved, or preventing the person perpetrating the abuse from coming within a certain distance of the premises, including 'ousting' a person from the premises. If multiple PPNs are issued for the same person this would be useful information for profiling a family to offer help and support especially in the situation where the victim is a pregnant women.

This work package aims to take the text components of the PPN to see if the following can be coded from the text: 1. Other agencies involved 2. Police action (police protection, arrest) 3. Referral to MARAC. This will enable possible follow up or action following the PPN to be coded and used to share with other agencies to examine outcomes following specific actions.

This work will use text mining to undertake identity extraction from text fields, such as the mention of social services, courts or health services (mental health or substance abuse services). Detect clustered terms that would suggest police action e.g. the term 'referred to' followed by any mention of the words police protection, police watch, court etc or the term 'arrest' which is preceded by 'under' 'following' or mention of known services such as 'Stay Safe'. This work will be conducted in partnership with South Wales Police to ensure appropriate words are mined.

Methods: Natural language processing would involve the development of a coding framework in the form of an ontology (e.g. concepts, such as words that are together for a specific meaning) which would be evaluated by manually coding a sample of text with words of interest and overall classification. This ontology (concepts) will be used to encode relevant features (e.g. any mention of the term e.g. “social services”). The frequency and the words surrounding the term (e.g. “referral to” “social services” (coded new social services), “previous” “social services” involvement (coded existing social services) will be utilised to automatically classify and identify the term in a coded form which can be quantified and integrated with other data. In summary, this package is using natural language processing to extract other organisations involved and police action taken from the NPP text data.

Work package 2: Scoping exercise of data available, system and quality. Qualitative assessment of systems currently in use and feasibility of harmonising all to use one system. Supervisor Professor Simon Moore

One of the barriers to this research includes the use of multiple software systems in the police. Babuta [18] discussed the issues with using multiple systems and reported the lack of consistency among data collection and reporting in the police force. Therefore, a scoping exercise would enable us to record what data is available on each system and what the similarities are to potentially harmonise the data into a minimum core dataset for sharing, or to identify what needs to be in place for each police force to use the same system. The purpose of the scoping exercise would be to map a range of opinions from individuals in the police, bringing together a variety of perspectives, a breadth of knowledge and expertise. This is aiming to consider where there are gaps or where novel approaches may lie to determine whether the barrier of multiple software systems can be overcome.

Methods: Members of police forces who input to the data sets, will be asked to take part in a short semi- structured interview about the software system their local police force uses to collect data and their opinions of this system and data quality (missing data) and what fields are collected. This would provide us with in depth knowledge and rich qualitative data regarding the opinions of various police forces about data collection and the systems they use. The interview would comprise of multiple set questions including open-ended questions to acquire detailed responses The interviews would include questions about the usability of the current system they use, the financial and other considerations in choosing the current system, and whether they would be willing and able to change the system they use, what questions they feel are core and appropriate for sharing in a minimum core set dataset and what are their barriers and potential solutions to sharing data. The questions to be assessed include the systems used and their quality. In this way, we aim to scope out what data is in all the branches and establish whether we can harmonise all the data and what needs to be in place in terms of social and financial considerations to facilitate data sharing. If so, comparisons could be made much easier if the data can be synthesised all on one system, even if it is an additional core dataset system. By the assessment of the systems currently in use, we can identify the feasibility of harmonising all forces to use one shared system. It is known that a number of police forces are continuing to work collaboratively on such as on the NICHE collaborative arrangements with other NICHE forces in the UK. The scoping study in this research may be able to effect decisions made within these collaborative arrangements around recording of data set across a number of different forces.

In summary, this work package will develop a scoping document of what data systems are used, the data fields, data quality, and user assessment of the system.

Work package 3: Exemplar case study. Supervisor Professor Mark Bellis

The work package has been developed to identify the vulnerable pregnant woman and the main risk factors associated with the domestic abuse/violence during pregnancy with the help of linked routine data. The aim of the study is to build a framework for early identification of the at-risk women. This research work will help the police, public health, and the social care organisations to build appropriate support and early intervention programmes required to the vulnerable families. This work focuses to be a policy-relevant research for the benefit of society. The evidence of the risk profile of pregnant women would support informed policy and decision making.

Currently SAIL repository holds the datasets ranges from primary care - Welsh Longitudinal General Practice (WLGP), hospital admission record - Patient Episode Database for Wales (PEDW), Emergency Department Data Set (EDDS), mental health, social services, Education data on schools and pupils, Lifelong Learning Wales Record, Maternity Indicators Dataset (LLWR), National Community Child Health database (NCCHD), Substance Misuse Dataset (SMD), Annual District Birth/Death Extract, Welsh Demographic Service Dataset (WDS). In recent times SAIL has got the agreement to hold the COVID-19 data such COVID-19 Test, Trace and Protect. With the help of anonymised data linkage across the wide variety of data sources we would develop a risk profile of individual in the secured platform and follow up the individual longitudinally.

This project aims to build a cohort of pregnant women in Wales by linking Maternity Indicators Dataset and National Community Child Health Dataset with the help of ALF in SAIL databank. The objective of the project is to identify the women who were admitted to hospital (from PEDW) or attended A&E (from EDDS) for assault, any intentional harm or domestic abuse during pregnancy. The risk exposures such as household member and their demographic characteristics will be obtained from the various routinely collected datasets which would support to build the risk profile of the pregnant women. WDS provides an encrypted number of residential address of individual in Wales know as Residential Anonymised Linking Field (RALF) [19] which can link the individual with their household members, their demographic characteristics (gender, week of birth), household structure (number of household member, adults, children in the household). RALF will be linked with a Welsh Index of Multiple Deprivation (WIMD) score which indicates the overall derivation level of the area of residence. The SMD provides the information if the woman lived with any household member who had alcohol or any other drug related problem during pregnancy. Household member with serious mental illness, depression or any other mental health related issue will be obtained by linking WDS and WLGP datasets. Any health care visit/support (including physical and mental health) the woman had during pregnancy will be obtained from the WLGP, PEDW and EDDS. The Public Protection Notification (PPN) data from police force will be linked with the household of the pregnant women to get an idea of the previous call outs that might have been received from these families. This will provide the information on risk and vulnerability of the pregnant women. With the help of traditional statistical methods (regression) and using supervised machine learning classification algorithms (Decision Tree, Random forest, Naive Bayes classifier) the main risk factors associated with the domestic abuse of the pregnant woman and the relative weight of the importance of the risk factors would be achieved. This would inform the police, policy maker and other care provider to have early warning risk factors to enable early intervention and to build the necessary support plan.

Work package 4 – Data Linkage. Supervisor Professor Sinead Brophy

To examine the identifiers found in each dataset and examine the linkage quality and success when linking multiple dataset of an example set of data e.g. PPN's (individual police identifier), education (school identifier), health (NHS number), courts (name/address) from the South Wales region. This work will examine percentage match, match quality (fuzzy matching), and proportion of the dataset where matching across all 4 datasets is feasible. This work can be used to evidence base if name/address/dob are consistent and reliably found in different datasets. This will inform if these fields can be used to link data and what difference it would make if we could improve the quality of the collection of these fields. The findings from this work can be compared with the ideal situation in linking data, the use of a single person identifier as is found in Scandinavian countries in national identity numbers. The impact of this work is to inform the improvement and quality of data collection to improve data sharing in the future. Numerous health outcomes lie at the intersection of health and criminal justice, for example violence and substance abuse. Moreover, children developing in environments where guardians and parents engage in such behaviours are at risk of experiencing long- and short-term detriments to health. Opportunities for early intervention and mitigation of harm typically involves a multi-agency approach (e.g. MARAC). However, evidence to inform activity and evaluate the success of safeguarding (e.g. domestic abuse advocates) is limited. The impact of linkage is likely to be substantial in a UK context.

Dissemination, Outputs and anticipated impact :

Outputs– Peer reviewed publication on 1. national language processing for sharing police data, will it work? 2. linking multiagency routine datasets for preventing domestic abuse in pregnancy. Reports to the police and Welsh Government 1. scoping of datasets and recommendations for core data for national sharing, 2. quality of linkage and mechanisms to improve linkage.

A half day multiagency workshop will be delivered to disseminate findings and to discuss solutions and ways forward in sharing whole datasets nationally as well as locally.

What do you intend to produce from your research? This work is intended to be a step in the direction of bring police data from all of Wales into SAIL where it can be linked to education, CAFCAS (family court), medical records (GP/Hospital/A&E including mental health, medication, substance abuse, assault/intentional harm) all of which can be used to evaluate patterns and trends on a population level.

The scoping of datasets can be used to inform the best way of sharing entire police datasets and coming to a consensus on nationally agreed minimum dataset fields that could be shared nationally and combined with other agencies datasets.

The exemplar cases study is part of work with Public Health Wales which sets out to examine if individual profiling using linked routine data can be used to help target services. This case study will demonstrate the value of linking PPNs with health data in order to predict a A&E attendance for assault/abuse. If linked data can be used to profile the families where abuse escalates to an A&E attendance, and this can be also feedback to the police, future contacts with the family for all agencies can be based on a fuller picture.

What do you think the impact of your research will be and for whom? The linked police & SAIL dataset can be compared and contrasted with the police-MOJ data in England [20]. In the longer term this linkage will enable research across England, Wales and Scotland to better understand crime in terms of population trends, compare and contrast policy decisions in the different nations and evaluate impact in a natural experiment design.

In the short term, this work helps to promote, and evidence based the best way of unlocking police data for public health policy and practice.

Time table:

April – June: WP1: Acquisition of PPN data and working with South Wales to define the NLP. WP2, Ethical approval, organisation of interview times for scoping of datasets. WP3: defining the variables for data extraction.WP4: Examination of linkage quality.

July – Sep: WP1: NLP analysis, WP2, Scoping of datasets and writing up of scoping recommendations and discussion on possible minimum dataset. WP3: analysis of NNP/MIDS/A&E/GP/NCCHD/hospital datasets for predicting abuse.

Oct – Dec: Writing up of findings, finalise reports, papers and outputs. Multi-agency Workshop to discuss next steps in light of findings.

Project management costs and governance: The project team will meet using zoom/remote methods every month. All day to day management and governance will be through the project/co-applicant team.

Ethics and regulatory approvals: This study will seek university ethical approval to undertake interviews with the police. This approval can take about 2-4 weeks so it is anticipated that it will be in place long before the start of this work. The IGRP approval for linking data is already in place. The data sharing agreements with South Wales Police are in place and discussions are already underway with North Wales and Mid-Wales police forces.

PPI : This study will work with the existing patient and public involvement group of the National Centre for Population Health and Research and will working with the SAIL Consumer panel who have a great deal of experience in linked data research and datasets for public benefit.

Success Criteria and barriers to proposed work: This work will be seen as a success if we (1) acquire PPN data and link it to health and education data, (2) agree a way forward in harmonising one type of data (e.g. PPN) for Wales which can be shared in an accessible linkable and secure way (e.g. within SAIL) (3) can definitively state if text mining could be taken forward to enhance data sharing or if another strategy is needed (4) have a report on best methods of linking the same individual across datasets.

The main barrier to the work is the changing situation with COVID 19 in terms of priorities for Public Health Wales, the Police and universities. If the COVID situation worsens it may happen that co-applicants are pulled to do more rapid response and front line COVID work (Public Health Wales and Police) and university employees are pulled to do more COVID rapid response research analysis and online teaching/small group face-to-face work (e.g. practical’s that were delivered once in larger groups now need to be broken into multiple small socially distance groups and repeated multiple times). If this situation happened it is anticipated the work would continue but maybe at a slower pace than originally anticipated (e.g. setting up interviews may take longer, acquiring data may be slower as SAIL prioritise analysis for Welsh Government Technical Advisory work).