

Trying to see the bigger picture: a review of the opportunities for linking eHealth and social science datasets to enhance understanding of risk of suicide in Scotland

Alison Dawson^{1,2} and Margaret Maxwell^{1,2}

¹Data Management through e-Social Science (DAMES), an NCeSS research node

²Department of Applied Social Science, University of Stirling

Email address of corresponding author: margaret.maxwell@stir.ac.uk

Abstract. In this paper we make the case for linking the Scottish Morbidity Record (SMR) with Scottish Census data to produce a linked dataset that would allow analysis of the interplay between known risk and preventative factors for suicide and self harm. We map those factors to data contained in the two datasets. We briefly discuss methods of data linkage, ethical objections to linking data, and technical difficulties around linkage, before providing examples of data linkage projects to support the feasibility of our proposals.

Introduction

We already know much about trends in suicide in Scotland. In 2007 there were 838 deaths by suicide in Scotland (defined as deaths from intentional self harm and events of undetermined intent), an age-standardised rate of 15.9 per 100,000 population per year (ISD Scotland 2008). Suicide rates are three times as high for men as for women in Scotland, and rates of suicide in the most deprived areas of Scotland were double the Scottish average (ibid). We know that for men the highest rate of suicide occurs in the 25-34 age group, and for women in the 45-54 age group (Platt et al 2007). We know that the most common method of suicide amongst males is hanging and amongst females is self-poisoning/overdose, although there are variations to this pattern - for instance, drowning is the most frequently used method of suicide for males in the Highlands and Islands (Platt et al 2007), and death by firearms is over-represented in suicides of farmers and farm workers compared to the Scottish average rate for this method (Stark et al 2006).

These kinds of statistic highlight particular risk factors associated with suicide. To date, the study of risk and protective factors for suicide has tended to result in studies of single risk/protective factors which focus on either illness or psychological or social factors (such as previous mental illness, coping behaviours, unemployment, religious beliefs, deprivation, population density etc.). Fewer attempts have been made to (statistically) model the interplay between different risk and preventative factors in populations completing suicides (McLean et al 2008; for an exception see Hawton et al 1993). We could model the interplay between

previously identified risk factors in a Scottish population if we had access to a dataset that contained all the pertinent health, social and economic variables for individuals whose deaths were recorded as resulting from intentional self harm or events of undetermined intent. Such a dataset does not currently exist, but there is the potential to create one.

The potential benefits of record linkage for health-related research have been recognised for some time (Newcombe et al 1959; Newcombe 1988). Probabilistic record linkage methodologies have similarly been developed and refined over an extended period (E.g. Fellegi and Sunter 1969; Kendrick and Clarke 1993). But in the past high computing resource requirements and technical difficulties have limited its use. eResearch tools such as GRID computing technologies, and developments such as the availability of public domain software for record linkage, are now opening up this avenue of research to a far wider research community. Within the Data Management for e-Social Science (DAMES) project (<http://www.dames.org.uk/>) we are exploring the potential to link Scottish and UK eHealth data resources with pertinent social science databases. This will produce a powerful tool with which to address this deficit and enhance our understanding of risk of suicide in Scotland.

Creating a more complete dataset

Relevant data on individuals who have attempted or completed suicides is spread across a number of datasets owned and administered by different parties. The General Register Office for Scotland (GROS) collects data on cause of death. This data is linked by ISD Scotland to hospital discharges from non-obstetric specialties, psychiatric inpatient information and cancer registration data contained in the Scottish Morbidity Record (SMR). Requests can be made to access specific data contained in the combined dataset, known as the Linked Acute Dataset. Thus, a population of people having completed suicides can be identified and certain of their healthcare episode data examined. Those people who have attempted suicide may be identifiable through examination of the SMR, in particular 'SMR01', the scheme that relates to general/acute inpatient and day case episodes. The populations thus identified may have had prior contact with primary healthcare services (e.g. General Practitioners) or with secondary healthcare services (e.g. hospitals, specialist clinics) on an inpatient or outpatient basis. Access to the primary healthcare records of the population of interest is problematic, since there is no centralized resource for these. Scottish General Practices individually control access to the primary care data that they collect, although some practices may contribute data to collective resources such as the Practice Team Information (PTI) scheme run by ISD Scotland, which covers approximately 8% of the estimated Scottish population. There is however a centralized resource for secondary healthcare episodes not contained in the Linked Acute Dataset. SMR data are arranged in 'schemes', each related to a different type of secondary healthcare system interaction. Once an individual is identified, it is possible to obtain details of their healthcare interactions from across all of the SMR schemes.

Whilst the SMR can provide a great deal of health-related data, it tells us very little about risk and preventative factors that depend on the wider social and economic circumstances of the individual. For this we need social survey data. Many large scale social surveys are based on samples of the population, and as a result are likely to include relatively low numbers of people who have attempted or who have subsequently completed suicides. Although only collected once every ten years, the Scottish Census, administered by the Office for National Statistics (ONS) and GROS, samples the whole population and creates a 'snapshot' of the person with data on household living arrangements, family circumstances, education, (un)employment, religious beliefs, and geographical mobility.

Table I. Data relating to suicide risk and preventative factors contained in selected datasets.

	SMR	Census
Known risk factors (increase the risk of suicide and self harm unless otherwise stated)		
Diagnoses of mental illnesses and/or a general history of psychiatric treatment.	Yes	
Prior history of self harm and attempted suicide.	Yes	
History of substance misuse – highest risk with opiate use disorders.	Yes	
Epilepsy – increase in risk varies by type and severity.	Yes	
Personality traits such as hopelessness; neuroticism; extroversion; impulsivity; aggression; anger; irritability; hostility; anxiety; low problem-solving skills.	Yes	
Episodes of PMS and points in menstrual cycle with low oestrogen.		
Pregnancy	Yes	
Abortion, compared to carrying to full term (but there may be confounding factors involved).	Yes	
Unemployment		Yes
Occupational social class (generally inverse relationship to suicide and deliberate self harm, with higher risk for those in lower classes)		Yes
Living in areas with greater levels of socio-economic disadvantage.		Yes
Having been subject to sexual abuse.	Yes	
Known preventative factors (reduce the risk of suicide and self harm unless otherwise stated)		
Coping skills, including; problem-solving skills; skills involving self agency; and control of emotions / thoughts / behavior.		
Reasons for living – high levels of future orientation and optimism, resilience factors.		
Physical activity and health – participation in and positive attitude to sport; perceptions of positive health.		Yes
Family connectedness, including; good relationships with parents for adolescents; having children living at home for women.		Yes
Supportive schools, including having access to healthcare professionals for adolescents, especially those who have experienced sexual abuse, have learning difficulties, or identify as LGBT.		
Social support		
Religious participation, depending on levels of secularization and social and cultural integration.		Yes
Employment, especially full time.		Yes
Exposure to suicidal behavior, either through media accounts or through friends and relatives.		
Social values – traditional social values for adolescent girls, individualistic values for adolescent boys.		
Access to health treatment		

Mapping known risk factors to available data

It is important when requesting data to ensure that what is asked for is adequate, relevant and not excessive. The datasets that we are interested in linking contain considerable data that is not of interest to a research study with the goal of enhancing understanding of risk of suicide in Scotland. Accordingly, one of our first tasks is identify which variables in the datasets in which we are interested (SMR, Census) have the potential to shed light on one or more of the known risk factors and protective factors against suicide identified in McLean et al's (2008) systematic international literature review of review-level data. Table I, above, summarises

this mapping exercise at the level of the dataset, indicating where we believe each will be able to complement the other and add to the power of the linked dataset for analyses of risk of suicide. As can be seen from Table I, the linked dataset will considerably expand the number of factors that can be examined, but there are still some that will not be addressed without further data linkage. A more detailed account of this mapping exercise, which identifies and maps individual variables or groups of variables to both known risk and preventative factors and to suspected risk and preventative factors for which there are gaps in the review-level evidence (also identified by McLean et al 2008) can be found on the DAMES website.

Linking datasets: methods

There are three main types of data linkage technique: match-merge, deterministic and probabilistic (Campbell 2009). The match-merge method requires the presence in both datasets of a single common identifier, such as the Community Health Index (CHI) number – a number unique to individuals on a Scottish population register used for health purposes. This method would be inappropriate for the research suggested here because there is no single common identifier to link SMR and Census data. Deterministic record linkage also requires an exact match of identifying information but uses multiple criteria to establish a match. So, for example, the rule might be to link data for cases where surname, date of birth and postcode were all identical across both datasets. With deterministic methods there is a binary outcome – cases are either linked or not linked. Deterministic methods are considered less well suited to larger datasets. Probabilistic record linkage methods use algorithms to ‘score’ pairs of cases in terms of the likelihood that they relate to the same individual by comparing data on a range of different person-specific variables, with linkage then being made when the calculated statistical probability of a match exceeds a certain threshold. Probabilistic methods are often preferred as they consider all the available linking variables but do not require exact matches. Both deterministic and probabilistic methods typically incorporate algorithms to help avoid false negatives, i.e. failing to link data on the same individual across datasets due to variation in the matching variable data. For example, phonetic algorithms such as Soundex, an algorithm for indexing names by sound as pronounced in English, can be used on name data to help avoid this problem as a result of different spellings.

Issues and solutions

There are a number of issues around data linkage which may act as barriers to the successful completion of the proposed project. Chief amongst these are ethical issues around the use and linking of personal data and technical difficulties with data linkage, both as a research method and due to the infrastructure which may be required in order to address security and confidentiality concerns.

Ethical objections to data linkage revolve around issues of: informed consent; confidentiality; data security; and disclosure. It has been argued (Boyd 2007) that if individuals are not told about the possibility of subsequent data linking, then their consent to providing data was not fully informed and their data cannot ethically be used for additional linking studies until informed consent for the new use has been sought and obtained. This would be the case in relation to the Census and SMR data that we plan to use. However, there is support for the contention that the need to seek specific consent should be waived in circumstances where there is a clear public interest in overriding consent requirements *and* measures are in place to ensure that data is effectively anonymous to those with no need to know individual identities (Manson and O’Neil 2007; Muir 2007). Anderson (2008) expresses concerns that increasing

linkage of health data will encourage and make easier both unlawful access to health records, such as by those pretending to have appropriate NHS status, and lawful access, for example by agencies such as the police, to confidential health data that discloses unlawful activities such as substance misuse or under-age sexual intercourse. We and others involved in data linkage argue that these concerns can be adequately addressed by building appropriate safeguards, including anonymisation and encryption, into data linkage systems (e.g. Sinnott et al 2006; Trutwein, Holman and Rosman 2006; Lyons et al 2009). Disclosure relates to concerns that a third party with access to anonymised datasets can identify individuals or, knowing that the dataset includes particular individuals, can identify the attributes of those individuals (Gutmann et al 2008). There is a tension between minimizing the risks of disclosure and maximizing the usefulness of linked data, but again we would argue that adequate safeguards can be built into infrastructures for data linkage.

In terms of technical difficulties, we already know that it is possible to link certain SMR and Census data. The Scottish Longitudinal Study (SLS) has already successfully linked census, health, and vital events data for a semi-random sample of 5.3% of the Scottish population (Hattersley and Boyle 2007). The SLS sample is not sufficiently targeted for our purposes, and does not include all of the data that we wish to link, but provides us with a model for the linkage process. We also already know that it is possible to design systems for linking data and providing access to the resultant datasets that satisfy the security and confidentiality requirements of data owners whilst providing research access to linked data. For example, the Western Australia Data Linkage System (WADLS), established in 1995, uses probabilistic matching to link data from more than 30 administrative and health datasets and unique encrypted identifiers to replace personal data so that requested linked data may be provided without personally identifying information. (Trutwein, Holman and Rosman 2006). Similarly the Secure Anonymised Information Linkage (SAIL) databank being developed in Wales uses a system in which resource providers split the personal and clinical data at source, forwarding the former to a 'trusted third party' to allocate a unique 10-digit number which replaces the personal information and allows the clinical and anonymised personal information to be rejoined (Lyons et al 2009). We are developing a system that shares some of these features. Our intention within DAMES is to use our research on suicide and self harm as a vehicle to help develop e-infrastructures that will enable the linking of health and social science datasets more generally. As part of the DAMES project, colleagues based at the National eScience Centre (NeSC) in Glasgow have developed e-infrastructures that exploit the results of previous projects (such as the Virtual Organisations for Trials and Epidemiological Studies (VOTES) project) to enable secure data anonymisation and to support secure, federated access to a variety of anonymised distributed and heterogeneous data sets (See Sinnott et al 2009 for technical details). These systems are now available for demonstration with sample data.

Conclusions

In this paper we have argued that a linked dataset is necessary in order to model the interplay between previously identified risk and preventative factors for suicide and self harm in a Scottish population. We have selected the Scottish Morbidity Record (SMR) and the Scottish Census 2001 and we have mapped the variables within these datasets that link to known risk and preventative factors. We have shown by examples that this linkage is possible and that systems can be constructed that address ethical concerns around data linkage. We have developed demonstrators for e-infrastructures that will provide anonymised linked data in secure environments with federated access arrangements. Our next step is to formally approach the data owners and to move this project beyond the demonstration phase.

Acknowledgments

We gratefully acknowledge the role of the Economic and Social Research Council (ESRC) in funding the DAMES project, via the National Centre for e-Social Science (NCeSS).

References

- Anderson, R. (2008): 'Patient confidentiality and central databases', *British Journal of General Practice*, Vol. 58, No. 547, pp. 75-76.
- Boyd, K. (2007): 'Ethnicity and the ethics of data linkage', *BMC Public Health*, Vol. 7, p318.
- Campbell, K.M. (2009): 'Impact of record-linkage methodology on performance indicators and multivariate relationships', *Journal of Substance Abuse Treatment*, Vol. 36, pp 110-117.
- Fellegi, I.P. and Sunter, A.B. (1969): 'A theory for record linkage', *Journal of the American Statistical Association*, Vol. 64, pp. 1183-1210.
- Gutmann, M.P., Witkowski, K., Colyer, C., MacFarland O'Rourke, J. and McNally, J. (2008): 'Providing spatial data for secondary analysis: issues and current practices relating to confidentiality', *Population Research and Policy Review*, Vol. 27, pp. 639-665.
- Hattersley, L. and Boyle, P. (2007): 'The Scottish Longitudinal Study: An introduction', LSCS Working Paper 1.0. Available at <http://www.lscs.ac.uk/sls/publications.htm>.
- Hawton, K., Zahl, D. and Weatherall, R. (2003): 'Suicide following deliberate self-harm: long term follow-up of patients who presented to a general hospital', *British Journal of Psychiatry*, Vol. 182, pp. 537-542.
- ISD Scotland (2008): '*Suicide statistics 2007*', Statistical Publication Notice dated 26 August 2008. Accessed at <http://www.isdscotland.org/isd/5724.html>.
- Kendrick, S.W. and Clarke, J.A. (1993): 'The Scottish record linkage system', *Health Bulletin (Edinburgh)*, Vol. 51, pp. 72-79.
- Lyons, R.A., Jones, K.H., John, G., Brooks, C.J., Verplandke, J.P., Ford, D.V., Brown, G. and Leake, K. (2009): 'The SAIL databank: linking multiple health and social care datasets', *BMC Medical Informatics and Decision Making*, Vol. 9, pp. 3.
- Manson, N.C. and O'Neill, O. (2007): '*Rethinking Informed Consent in Bioethics*', Cambridge University Press, Cambridge.
- McLean, J., Maxwell, M., Platt, S., Harris, F. and Jepson, R. (2008): '*A systematic international literature review of review-level data on suicide risk factors and primary evidence of protective factors against suicidè*', Scottish Government, Edinburgh. At <http://www.scotland.gov.uk/Publications/2008/11/28141444/0>.
- Muir, R. (2007): '*eHealth, secondary uses and information governance in NHS Scotland. A discussion paper*', Available at http://www.isdscotland.org/isd/servlet/FileBuffer?namedFile=eHealth_secondary%20uses.pdf&pContentDispositionType=inline.
- Newcombe, H.B. (1988): '*Handbook of record linkage*', Oxford University Press, Oxford.
- Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P. (1959): 'Automatic linkage of vital records', *Science*, Vol. 150, pp. 954-959.
- Platt, S., Boyle, P., Crombie, I., Feng, Z. and Exeter, D. (2007) '*The epidemiology of suicide in Scotland 1989-2004: an examination of temporal trends and risk factors at national and local levels*', Scottish Executive, Edinburgh.
- Sinnott, R.O., Stell, A.J. and Ajayi, O. (2006): 'Initial experiences in developing e-health solutions across Scotland', Workshop on Integrated Health Records: Practice and Technology, Edinburgh, March 2006.
- Stark, C., Gibbs, D., Hopkins, P., Belbin, A., Hay, A. and Selvaraj, S. (2006): 'Suicide in farmers in Scotland', *Rural and Remote Health*, Vol. 6, pp. 509-518.
- Trutwein, B., Holman, C.D.J. and Rosman, D.L. (2006): 'Health data linkage conserves privacy in a research-rich environment', *Annals of Epidemiology*, Vol. 16, No. 4, pp. 279-280.