

Linking social care, housing & health data

Data Linkage literature review 2010

Paper 1

What is data linkage?

Data linkage or record linkage is when information from two or more records of independent sources are brought together, when they are perceived to belong to the same individual, family, event or place (Brook et al. 2008:19). It is the technique that is used to connect the information across several disparate data sources (this is most commonly used currently within medical and population health research) (Karmel & Rosman 2008). According to Flowers & Ferguson (2010: 1) with regards to health information ‘people who access and use health intelligence need ‘one-stop shops’ of information, removing the need to hunt around in several places for different pieces of a ‘health intelligence jigsaw’’. Data linkage allows different types of information to be more readily available and so reduces the length of time looking for data.

According to Holman (2009) there were only a few places internationally that have linked data within large, multiple, population-based administrative data sets. The main examples of these include the: Manitoba Population Health Information System; Oxford Record Linkage Study; Scottish Record Linkage System; Rochester Epidemiology Project; Centre for Health Services and Policy Research in British Columbia (Canada); Western Australia Data Linkage System (WADLS) and the Centre for Health Record Linkage (CheReL) in New South Wales (Australia). See Table 1 for further information of these examples. Having looked at these examples and searched the available literature; the benefits of data linkage have been identified and the next section will explore some of these benefits. There will also be a discussion of some examples of good practice with regards to data linkage, access to data and potential limitations that need to be considered.

Table 1: Worldwide Data Linkage Systems

<p>Canada</p>	<p>British Columbia Linked Health Database – this was established in 1996 and covers datasets on the medical services plan, PharmaCare, hospital separations, continuing care, birth registrations, death registrations, mental health episodes records, workers compensation board and the British Columbia Cancer agency cancer incidence file.</p> <p>Manitoba Population Health Information System – since 1970 the provincial health department has provided anonymous health care files into a repository; these are then linkable with a scrambled identifier which is unique to a family (not the individual). For more information see: http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?definitionID=103348</p>
<p>USA</p>	<p>Rochester Epidemiology Project – this contains medical records since 1935 on all residents of Olmsted County, Minnesota in a single research database. Records are tracked using a computerised bar coding system. See http://www.rochesterproject.org/ for more information.</p>
<p>UK</p>	<p>Oxford Record Linkage Study (ORLS) – first data linkage system developed in 1962. It contains computerised records (around 10 million records) for 5 million people in the Oxford area, the ORLS contains irreversibly anonymised abstracts of records of morbidity, births and deaths from 1963-1999. following 1999, the data collection has discontinued and all patient identifiers have been removed from data sets.</p> <p>Scottish Record Linkage System – linked datasets include morbidity, maternity, neonatal and mental health records, cancer notifications and ambulance and emergency centre attendances. In 1992 a Scottish central population index issues a single, unique NHS identification number to each individual within the population of Scotland which can then be used for matching records.</p> <p>Work & Pensions Longitudinal Study (WPLS) – linked benefit data from the Department for Work & Pensions (DWP) and employment data from Her Majesty’s Revenue & Customs (HMRC). Established in 2004 to provide statistics, but is also used for a limited number of operational purposes (investigation of fraud).</p>
<p>Australia</p>	<p>Western Australian Data Linkage System – this is a multi-set system which links health and welfare related data. It uses computerised probabilistic matching and clerical review to create a master linkage key between the records. For more information see: http://www.datalinkage-wa.org.au/</p> <p>NSW Centre for Health record Linkage – this was established in 2006 to create and maintain a record linkage system for health and human services in NSW and the ACT (Australian Capital Territory). This uses probabilistic record linking techniques to enable research, planning and evaluation for populations within the covered area. See http://www.cherel.org.au/ for more detail.</p>

<p>*Developing or incomplete systems</p>	<p>Icelandic Health Sector Database New Zealand Encrypted National Health Index Swedish Demographic Database</p>

Benefits of linking data

Cost: Once the data linkages have been made and then preserved, it means that the cost of linkage does not have to be incurred again for each new project that requires linked data (Chamberlayne et al. 1998) and so is more cost effective in the long run. Gisser & Haukka (2004: 113) believe that the use of ‘existing administrative data in research is attractive, since the total study costs...can be reduced significantly’. In a time of concern over spending, the cost of public services and the drive for efficiency the current available data is (and will continue to be) essential in ensuring population health and quality of care is maintained and improved whilst ensuring costs do not rise. Linked data will perhaps be seen as an essential way of providing such information and could be a more cost efficient way of using readily available data to inform policy; and it will minimise costs for new primary research (Flowers & Ferguson 2010; Christen & Goiser 2007).

Specifically looking at the example of WADLS, it is suggested that the cost of the data linkage system over the past 10 years has been relatively inexpensive compare dto the biomedical research infrastructure (Holman 2009), demonstrating this investment is cost effective in the long run. Linking existing data has been found to be cheaper and more effective than undertaking longitudinal studies and other more traditional research approaches within the health services (this is especially the case when trying to target more vulnerable or mobile groups within society). For the WADLS project, due to the lowered cost, it has meant that more research can be undertaken with the limited resources available.

Time benefits: Once linkages are made then this will reduce the length of time take to carry out projects and also may mean projects are more feasible. The time on primary data collection is significantly reduced, and may not be necessary at all (Gissler & Haukka 2004).

Improved data quality and integrity: linking data will improve data quality and integrity of the data already being collected (Christen & Goiser 2007). This is perhaps because people now see the data as being of more benefit in the longer term and so which to ensure that it can be used in the future. Also, the data linkage has identified and removed many duplication errors and other technical glitches which otherwise would not have necessarily been picked up, this has led to greater accuracy of recording data at the administrative level (Holman 2009).

Making better use of available data: data linkage allows better use to be made of already existing information, the data can be re-used repeatedly for a variety of new studies (Christen & Goiser 2007). It is also anticipated that if primary research (data collection) is ongoing, then data is being added to the already collected and stored data and it will be possible to create an even more sophisticated understanding of what is occurring and how perhaps issues might change over time (Hanlon et al. 2007).

Privacy and consent: privacy and confidentiality has been raised as a concern for data linked sets, however, a surprising outcome from the data linkage is that it has actually conserved patient privacy (Holman 2009). This is because it reduces the need for names and other personal identifiers to be made available to researchers (previously this information had been required by researchers in order for them to be able to find and clerically gather information on individuals).

Communication benefits: data linkage systems mean that there has to be an increase in communication between researchers, clinicians, administrators, consumer groups and the media as was demonstrated within WADLS (Holman 2009). It increases team working, allows for improved cooperation and identification of future possibilities and allows for debate about the uses of data and results of any subsequent research.

Research benefits: data linkage can be available for researchers in the academic environment, industry, government organisations and the wider community. It can allow for a wide range of projects to be undertaken with the data, and allows for many different projects at a lower cost compared to undertaking primary research as individuals/groups (Holman et al. 2008). The appropriate use of existing population-

based data collections is acknowledged to be an efficient and effective alternative to individual longitudinal field studies (Sibthorpe et al. 1995; Brook et al. 2008). Also, the linking of such resources means that monitoring, surveillance and analytical assessment of the total population can occur. Hanlon et al. (2007: 405) believe that researchers should take advantage of such a resource and it will allow for areas of relative 'ignorance' to be addressed, if the linked data was not available then the research would not be possible.

The linking of data also allows for the evaluation of service outcomes and achieving best clinical practice has been demonstrated (Semmens et al. 2001; Kelman et al. 2002; Hanlon et al. 2007). Some other areas where linkages could be made have been identified within WADLS, these include: cross-jurisdictional linkage (e.g. collaboration between different agencies, both government and non government); genealogical linkage (this was where they aimed to identify family links of all nuclear families so this could be used in combination with health data and allow for population based genetic and human genome research); links to state electoral roll which would allow for longitudinal studies of outward migration which can then link to health records; geocoding of residential addresses allowing for links to environmental factors and to improve the identification of social disadvantage and remoteness indices.

Commercial benefits: using WADLS as an example, this data linkage system is reported to have had some commercial and competitive benefits for Western Australia. This is because it has provided the research community within Western Australia (WA) with a competitive advantage in certain situations, compared to those areas which do not have a data linkage system. This has meant more research income has been obtained from sources outside of WA, that might not have otherwise materialised. Due to this, Holman (2009) suggests the revenue has increased employment and helped to stimulate the local economy. It has also attracted and retained high quality researchers which might not have otherwise come to the area, and finally it has had benefits for policy development and for improvements in quality of the local health system. Linked to this are then improvements in the population's health. This is because there has been an increase in medical and scientific knowledge from the research which can then help improve

services and care and also it can help lead to reforms in policy and legislation (in the case of WA, reforms in mental health legislation and service delivery can be attributed to WA data linkage-based research).

How it can be used to support policy

According to Christen & Goiser (2007) linked data could contain information that is needed to improve health policies and could also extend to other policy areas if the linked data is available (Kelman et al, 2002; Bass & Carfield 2002). As can be seen within the previous section there are many benefits of data linkage, it can allow research projects to be undertaken that might not have been feasible previously. In linking health and social care it might be possible to determine how these link and the outcome it has for thus helping to inform policy for the benefit of those using the services.

When looking for specific benefits, using WADLS for an example, this project has allowed for significant reforms within health policy and improvements in clinical practices throughout WA, thus showing the benefit of good health data linkage and the research which arises from this (Brook et al. 2008). The linkages of data means that research can be undertaken in to all major diseases, their risk and protective factors and also how health services are used and the outcomes that are achieved (Holman et al 2008). Further to this the benefits of linking different types of data (such as health and social care data) can be undertaken and used to inform and develop policy.

Kemm et al (2010) discuss specifically the benefit of social care and health data linkage to not only support policy, but also in improving data quality. In England (and similarly to Scotland), there are several separate social care statistical data collections and publications. However, “although these publications appear to contain a great deal of information, in reality they give a very incomplete picture of clients receiving care and of the care that they receive. Using the published statistics, it is not possible to build a picture of the prevalence of ill health in social service clients, or the care pathways that they follow”. Data linkage could provide

evidence on the prevalence of conditions such as dementia, stroke, arthritis, diabetes and heart problems. “People with these conditions have little contact with secondary or even primary health care, but considerable contact with social services”, but this detailed information is not collected in social care statistics. Kemm et al (2010) go on to discuss how data linkage could be used to inform end of life care. Additionally, from the greater use of social care and health data, it “will encourage improved accuracy in the data recording and reporting systems”.

A further example of data linkage being instrumental in developing policy is that of the Work & Pensions Longitudinal Study (WPLS)¹. The WPLS was introduced in January 2004, and links benefit information held by DWP, with employment records from Her Majesty's Revenue & Customs (HMRC). New data-sharing provisions introduced in the Employment Act 2002 opened the way for further data sharing between DWP and HMRC.

The WPLS offers DWP the opportunity to significantly improve both the analytical evidence base and operational effectiveness of the Department. Its aims are focused around supporting DWP's agenda for Child Poverty, Welfare-to-Work and Retirement Income planning policy. This will assist DWP target their resources to the appropriate people, in the appropriate way. Information about DWP linking data for non statistical reasons can be found in the following documents (examples include the prevention of benefit fraud by sharing (Housing and Council Tax) data with local authorities):

www.dwp.gov.uk/docs/information-sharing.pdf

www.dwp.gov.uk/docs/dwp-your-personal-information.pdf

Best practice for linking data

Chamberlayne et al. (1998) suggest that ethical considerations must be made when undertaking data linkage, and providing this data for research. Care must be taken to ensure all relevant legislation is met and that the data stewards are aware of the

¹ http://research.dwp.gov.uk/asd/longitudinal_study/ic_longitudinal_study.asp

implications of linking the data which they are responsible for with other sources of data (Gissler & Haukka 2004). The introduction of a formal review process which is specifically designed to deal with data requests can help researchers to assess and monitor the requests being made to this resource. Access should be reviewed on a project by project basis by an ethical committee and also then separate approvals should be obtained from each of the agencies who are custodians of the data; then before analysts can receive the data, they should have to agree in writing to stringent set of conditions for data security (e.g. safe storage of the data, no transferring data to others etc.) (Holman et al. 2008). Gissler & Haukka (2004) also discussed how researchers should obtain permission to access the linked data. They suggest that it would normally be the institution who maintain the data which would have responsibility for granting access, however depending on the type of information authorisation may also have to be sought else where e.g. via a committee.

Within the WADLS project, a best practice protocol was developed which took into account privacy issues whilst allowing data custodians to retain control over the uses of the data within their care. For this data linkage, the following were considered and developed. If unique identifiers are available then the problem of linking data is significantly reduced as a simple database will be able to link all of the individual identifiers (Christen & Goiser 2007). However, if such an identifier is not available (which is often the case) then more sophisticated ways of matching the data will need to be employed.

Once personal identifiers are made/seen, then they should be kept separate from the actual data, and should also only be used at the initial data linkage stage. For this process all possible identifies should be used to produce a concordance file which can then later be used to extract the data required for the specific piece of research (Kelman et al. 2002). There needed to be a clearly defined approach for this which is collaborative and help develop a trusting atmosphere between data custodians. So the principles that should underly any protocol for inter-agency data linkage, according to Kelman et al. (2002: 252) are:

- *Maximise the protection of individual privacy;*
- *Provide linked data files only to nominated researchers involved in specific, approved research projects;*

- *Provide researchers with no more than the datasets required for their specific project; and*
- *Assure data custodians that those data which are their responsibility will be used appropriately and that security obligations will be met.*

With regards to the actual data linkage, this should be undertaken on an isolated and secure computer, and all copies of personal demographic data should be destroyed as soon as linkages have been made. Those individuals who are undertaking the linkage should not have any involvement with the analysis of the linked data and they should not have any communication with researchers at the unit record level (Kelman et al 2002).

For researchers themselves, they should be required to restrict access to the linked dataset, only those involved in the actual analysis should have access, and any results should be presented in a suitably aggregate form. Also all data custodians should be able to comment on any reports/articles that are produced from the linked data in the draft stages. Finally all coded data sets should be deleted once the analysis is complete and this should then be informed to the steering group. Within the Scottish Longitudinal Study (SLS), (this is a large-scale linkage study using data from a variety of sources including Census data, Vital Events data (births, deaths, marriages), National Health Service Central Register (NHSCR) data (migration in or out of Scotland) and NHS data (cancer registrations and hospital admissions) (Hattersley & Boyle 2007)) there are protocols in place similar to the WA. For example, there are straight controls in relation to the dataset itself as the SLS uses individual-level data for a sample of 20 birthdates, only a small number of researchers who have to maintain this dataset are allowed to know these 20 dates, also there is no name or address included within the dataset and the method used to flag and link data is a complex system specifically designed to maintain anonymity of the records (Hattersley & Boyle 2007).

Limitations of linking data

Team and collaborative working: the creation of a data linkage system is challenging and demands 'leadership, inter-agency and inter-sectoral cooperation and perseverance (Holman 2009: 9). It is suggested that the technological solutions and resources are not as important as the good relationship between institutional staff and leading individuals.

Privacy issues: there has been concern raised about privacy of data. It has been thought that data linkage could increase the likelihood of disclosure of individual and there have been considerable concerns of confidentiality issues and data ownership (Flowers & Ferguson 2010). However, has been seen within the case of WADLS, data linkage was actually found to reduce the risk of a breach of confidentiality (Holman et al. 2008).

Time and cost pressures: there is also an issue in relation to the speed which data linkage can occur (Flowers & Ferguson 2010), initially this can be a time consuming exercise and may not happen as rapidly as individuals might hope and initially there may be a high cost in undertaking this work, however, once the linkage has been undertaken then the cost and time for subsequent research projects are subsequently reduced (Gissler & Haukka 2004).

Appropriate variables/ data quality: Gissler & Haukka (2004) suggest that even if the data is of a good quality there could still be variance in the quality of different variables; for example: information of events that occur before or after the registered event on the data set (and so not reported on the data) means that information could be incomplete. It may also be the case that the data you are accessing may not have the specific or appropriate variables which you were wanting for an analysis. Griffith et al. (2009) found that many national datasets did not include factors at each level of the social-ecological framework and so they were not able to disentangle the factors they wished to undertake an analysis upon. They also found an issue with adequate representation of some variables; again if there is not adequate representation then an analysis will not be possible. There could also be an issue if the current information is not accurate, consistent or complete (Karmel & Rosman 2008). For example, if person-based linkages are over a long period of time, then

they will require access to a universal unique identifying number system or need extensive personal details to ensure accurate links are made. If this information is missing or incorrect/incomplete the linkage may not be possible.

Conclusions

As has been discussed data linkage has many benefits which outweigh the potential problems and limitations. Obviously, this is not an easy task to undertake and it is a complicated process, however there are useful examples which can be drawn upon that can help direct the linking of health and social care data within Scotland; so as to ensure that the data is used in the most efficient and effective way.

Flowers and Ferguson (2010) ask in their conclusion, 'what could 2020 look like?' they suggest that there will be no shortage of data, but unlike currently it will be more accessible in raw and aggregated forms which are easily linked with other available forms of data. This data will then be valued similarly to diagnostic tests and procedures to inform care and finally 'we will definitely be working more collaboratively and collectively for the good of our populations, and we will understand all patient data, properly used, are a public good to be cherished'. This is a hopefully and optimistic stance in which good and appropriate data linkage of high quality data will help achieve.

Louise Taylor (main author)

Ellen Lynch (contributor)

Health Analytical Services Division

Scottish Government (Spring/Summer 2010)

References

Bass, J. and Garfield, C. (2002) Statistical linkage keys: How effective are they? **Symposium on Health Data Linkage**. Available from:

<http://www.publichealth.gov.au/publications/symposium-on-health-data-linkage%3a-its-value-for-australian-health-policy-development-and-policy-relevant-research%3a-proceedings.html>.

Brook, E.L., Rosman, D.L. and Holman, C.D.J. (2008) Public good though data linkage: measuring research outputs from the Western Australian Data Linkage System. **Australian and New Zealand Journal of Public Health**. 32(1): 19-23.

Chamberlayne, R., Barer, M.L. and Lawrence, W.J. (1998) creating a Population-based Linked Health Database: A resource for Health Services Research. **Canadian Journal of Public Health**. 89(4): 270-273.

Christen, P. and Goiser, K. (2007) Quality and complexity measures for data linkage and deduplication. **Studies in Computational Intelligence**. 43: 127-151.

Flowers, J. and Ferguson, B. (2010) The future of health intelligence: challenges and opportunities. **Public Health**. Doi: 10.1016/j.puhe.2010.03.006 (article in press).

Frommer, M., Madronio, C., Kemp, S., Jenkin, R. and Reitano, R. (2007) **Population Health and Clinical Data linkage**. National Collaborative Research Infrastructure Strategy. Available from: <http://blogs.usyd.edu.au/ncrishealthdatalink/Issues%20Paper.pdf>. Date of access: 22/04/10.

Gissler, M. and Haukka, J. (2004) Finnish health and social welfare registers in epidemiological research. **Norsk Epidemiologi** 14(1): 113: 120.

Gray, L. Batty, G.D., Craig, P., Stewart, C., Whyte, B., Finlayson, A. and Leyland, A.H. (2010) Cohort Profile: the Scottish Health Surveys Cohort: linkage of study participants to routinely collected records of mortality, hospital discharge, cancer and offspring birth characteristics in three nationwide studies. **International Journal of Epidemiology**. 39: 345-350.

Griffith, D.M., Neighbors, H.W. and Johnson, J. (2009) Using national data sets to improve the health and mental health of Black Americans: challenges and opportunities. **Cultural Diversity and Ethnic Minority**. 15(1): 86-95.

Hanlon, P., Lawder, R., Elders, A., Clark, D., Walsh, D., Whyte, B. and Sutton, M. (2007) An analysis of the link between behavioural, biological and social risk factors and subsequent hospital admission in Scotland. **Journal of Public Health** 29(4): 405-412.

Hattersley, L. and Boyle, P. (2007) **The Scottish Longitudinal Study: An Introduction**. LSCS Working Paper 1. [online] <http://www.lscs.ac.uk/sls/LSCS%20WP%201.0.pdf>. (date of access: 04/05/10).

Holman, C.D.J., Bass, A.J., Rosman, D.L., Smith, M.B., Semmens, J.B., Glasson, E.J., Brook, E.L., Trutwein, B., Rouse, I.L., Watson, C.R., H de klerk, N. and Stanley F.J. (2008) A decade of data linkage in Western Australian: strategic design, applications and benefits of the WA data linkage system. **Australian Health Review**. 32(4): 766-777.

Karmel, R. and Rosman, D. (2008) Linkage of health and aged care service events: comparing linkage and event selection methods. **BMC Health Services Research**. 8(149). Doi: 10.1186/1472-6963-8-149. Available from: www.biomedcentral.com/1472-6963/8/149. Accessed: 18/05/10.

Kelman, C.W., Bass, A.J., and Holman, C.D.J. (2002) Research use of linked health data – a best practice protocol. **Australian and new Zealand Journal of Public Health**. 26(3): 251-255.

Kemm JR, et al.,(2010) Social care data in England: What they tell us and what they do not tell us, *Public Health*, doi:10.1016/j.puhe.2010.03.005

Lloyd, J.E.V. and Hertzman, C. (2009) From Kindergarten readiness to fourth-grade assessment: longitudinal analysis with linked population data. **Social Science and Medicine**. 68: 111-123

Lyons, R.A., Jones, K.H., John, G., Brooks, C.J., Verplancke, 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**. 9(3). Doi: 10.1186/1472-6947-9-3. Available from: www.biomedcentral.com/1472-6947/9/3. Accessed 14/05/10.

Semmens, J., Fletcher, D., Brameld, K. (2001) The role of epidemiology in achieving clinical best practice. **Cancer Forum** [Online] Available from: http://www.cancerforum.org.au/Issues/2001/July/Forum/The_role_of_epidemiology_in_achieving_clinical_best_practice.htm (date of access 30/04/10).

Sibthorpe, B., Kliwer, E., and Smith, L. (1995) Record linkage in Australian epidemiological research: health benefits, privacy safeguards and future potential. **Australian Journal of Public Health**. 19(3): 250-256.

Summary of articles and sources

Bass, J. and Garfield, C. (2002) Statistical linkage keys: How effective are they? **Symposium on Health Data Linkage**. Available from:

<http://www.publichealth.gov.au/publications/symposium-on-health-data-linkage%3a-its-value-for-australian-health-policy-development-and-policy-relevant-research%3a-proceedings.html>.

This article looks at existing measures of statistical linkage keys and has undertaken a project to assess if an analysis of data using deterministic mating of SLK's can lead to different results compared to using an analysis of 'real' linked data. The conclusion is that results are significant and therefore the method of the linkage must be considered when interpreting results of linked health data.

Brook, E.L., Rosman, D.L. and Holman, C.D.J. (2008) Public good though data linkage: measuring research outputs from the Western Australian Data Linkage System. **Australian and New Zealand Journal of Public Health**. 32(1): 19-23.

Article is reporting on the outputs that have occurred using WA data linkage system. Data on projects that were undertaken between 1995-2003; 258 projects were identified (specific examples given within article). Overall discussion highlighted that many of projects undertaken have resulted in reforms to health policy and improvement in clinical practices.

Chamberlayne, R., Barer, M.L. and Lawrence, W.J. (1998) creating a Population-based Linked Health Database: A resource for Health Services Research. **Canadian Journal of Public Health**. 89(4): 270-273.

This article suggests that once data linkage has been undertaken, then the cost will not have to be incurred again with further data linkage work on the same files. There is a focus on the policy and procedures that have developed regarding access to data, also ethical and confidentiality issues.

Christen, P. and Goiser, K. (2007) Quality and complexity measures for data linkage and deduplication. **Studies in Computational Intelligence**. 43: 127-151.

This article looks at the deduplication of a dataset or the linkage of several datasets. It discusses the methods that can be used to match records and how effective these might be. It gives an indication of how to assess the quality of the data linkage or deduplication.

Flowers, J. and Ferguson, B, (2010) The future of health intelligence: challenges and opportunities. **Public Health**. Doi: 10.1016/j.puhe.2010.03.006 (article in press).

Article looking at where information will be obtained and how research will be undertaken in the future (specifically relating to health information) and also what barriers are in place which will hinder access to the information. It focuses on data linkages and the potential benefits and problems of this in the future.

Frommer, M., Madronio, C., Kemp, S., Jenkin, R. and Reitano, R. (2007) **Population Health and Clinical Data linkage**. National Collaborative Research Infrastructure Strategy. Available from: <http://blogs.usyd.edu.au/ncrishealthdatalink/Issues%20Paper.pdf>. Date of access: 22/04/10.

This paper highlights practical issues for developing a linked data project. This generally focuses on the problems that arise on a national scale, in particular related to: legislation; policy; privacy & ethics; custodianship of the data; organisational capacity(in their ability to access the data); expertise (skills of individuals) and researcher engagement (how much does the researcher know in terms of legislation and policy issues linked to accessing data).

Gissler, M. and Haukka, J. (2004) Finnish health and social welfare registers in epidemiological research. **Norsk Epidemiologi** 14(1): 113: 120.

Article focuses on the registers of data held within Finland. It discusses unique identifiers and how to match up records. This article also has a discussion of data protection legislation in relation to these data registers and how data linkages will be affected by such legislation. Finally it discusses how researchers need to apply for permission for access to the data and what this process entails.

Gray, L. Batty, G.D., Craig, P., Stewart, C., Whyte, B., Finlayson, A. and Leyland, A.H. (2010) Cohort Profile: the Scottish Health Surveys Cohort: linkage of study participants to routinely collected records of mortality, hospital discharge, cancer and offspring birth characteristics in three nationwide studies. **International Journal of Epidemiology**. 39: 345-350.

Article looking at the results of linking SHeSs (Scottish health surveys). Currently there are few publications as the linkage for the most recent survey (2003) only occurred in 2007. Reports full potential of linking data has yet to be realised and there are many projects that can be undertaken once data is linked. Need to be aware of the potential issues relating to linked data and the information provided (accuracy etc.)

Griffith, D.M., Neighbors, H.W. and Johnson, J. (2009) Using national data sets to improve the health and mental health of Black Americans: challenges and opportunities. **Cultural Diversity and Ethnic Minority**. 15(1): 86-95.

This article explores how national datasets could be used to examine racial and ethnic disparities in health and mental health, if the data collected has the appropriate variables

Hanlon, P., Lawder, R., Elders, A., Clark, D., Walsh, D., Whyte, B. and Sutton, M. (2007) An analysis of the link between behavioural, biological and social risk factors and subsequent hospital admission in Scotland. **Journal of Public Health** 29(4): 405-412.

This article gives a specific example of research that has been undertaken within Scotland using linked data (1998 Scottish Health Survey and the Scottish hospital admission database). It highlights the benefits of linking national surveys.

Hattersley, L. and Boyle, P. (2007) **The Scottish Longitudinal Study: An Introduction.** LSCS Working Paper 1. [online] <http://www.lscs.ac.uk/sls/LSCS%20WP%201.0.pdf>. (data of access: 04/05/10).

This gives information about the Scottish longitudinal study, including its procedures and what data is linked from which sources.

Holman, C.D.J., Bass, A.J., Rosman, D.L., Smith, M.B., Semmens, J.B., Glasson, E.J., Brook, E.L., Trutwein, B., Rouse, I.L., Watson, C.R., H de klerk, N. and Stanley F.J. (2008) A decade of data linkage in Western Australian: strategic design, applications and benefits of the WA data linkage system. **Australian Health Review.** 32(4): 766-777.

This article looks at the development of data linkage within Western Australia, mapping its progress. It also discusses WADLS (Western Australia Data Linkage System) in detail and the benefit of this type of data linkage in detail.

Karmel, R. and Rosman, D. (2008) Linkage of health and aged care service events: comparing linkage and event selection methods. **BMC Health Services Research.** 8(149). Doi: 10.1186/1472-6963-8-149. Available from: www.biomedcentral.com/1472-6963/8/149. Accessed: 18/05/10.

This article looks at the process of event linkage and compared an event-based linkage method for identifying transition events between two care sectors in Australia with a long established longitudinal person-based linkage with uses helps with identification of event data for individuals.

Kelman, C.W., Bass, A.J., and Holman, C.D.J. (2002) Research use of linked health data – a best practice protocol. **Australian and new Zealand Journal of Public Health.** 26(3): 251-255.

This paper discusses a protocol for accessing administrative data when undertaking health service research. The aim of the protocol is to make data more easily available to researchers as it will allow for a controlled and secure access which protects the privacy of individuals data.

Lloyd, J.E.V. and Hertzman, C. (2009) From Kindergarten readiness to fourth-grade assessment: longitudinal analysis with linked population data. **Social Science and Medicine.** 68: 111-123

This article is an example of how data linkage has been used within a piece of research. It was looking at early child development within British Columbia.

Lyons, R.A., Jones, K.H., John, G., Brooks, C.J., Verplancke, 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.** 9(3). Doi: 10.1186/1472-6947-9-3. Available from: www.biomedcentral.com/1472-6947/9/3. Accessed 14/05/10.

This article looks at the SAIL databank within Wales and discusses the linking of datasets and the results from different ways of matching the records. This paper then concludes that once an infrastructure has been put in place and there is a reliable

matching process then records can consistently be found, The SAIL databank now offers a research ready platform for record linkage study and it can link both health and social care datasets.

Semmens, J., Fletcher, D., Brameld, K. (2001) The role of epidemiology in achieving clinical best practice. **Cancer Forum** [Online] Available from: http://www.cancerforum.org.au/Issues/2001/July/Forum/The_role_of_epidemiology_in_achieving_clinical_best_practice.htm (date of access 30/04/10).

This is a discussion of best clinical practice, and draws upon WADLS as an example of how data linkage can assist in achieving best practice.

Sibthorpe, B., Kliwer, E., and Smith, L. (1995) Record linkage in Australian epidemiological research: health benefits, privacy safeguards and future potential. **Australian Journal of Public Health**. 19(3): 250-256.

This paper discusses the importance of data, but raises issues about how collecting information from respondents can be expensive and often inefficient way of obtaining the longitudinal data necessary to study health effects. Therefore the paper discusses that the linking existing data can be an effective and efficient alternative. It also raises issues over consent and protection of privacy. It discusses these issues in relation to involving record linkage in Australia. The future potential for studies based on record linkage is discussed in the context of recent national and international developments in data protection

Further Resources (useful websites):

Website: <http://www.ihdln.org/member/#>

This provides information on the International Health Data Linkage Network and also its members.

Website: <http://www.publichealth.gov.au/publications/symposium-on-health-data-linkage%3a-its-value-for-australian-health-policy-development-and-policy-relevant-research%3a-proceedings.html>.

This provides a link to the papers presented at the 2002 Symposium on Health Data Linkage in Sydney, Australia in March 2002.