Version: 8th March 2017 (HT)

BRIDGEHEALTH
BRidging Information and Data Generation for Evidence-based Health Policy and Research

HA4. Exchange of health information

HA Leader: Hanna Tolonen

National Institute for Health and Welfare, Department of Health, Health monitoring Unit, PO Box 30, 00271 Helsinki, Finland
e-mail: hanna.tolonen@thl.fi, tel: +358 29 524 8638, Skype: hktolonen

This project is funded by the Health Programme of the European Union

Table of Contents

Working group (contributors marked with bold)

Angela Fehr (FehrA@rki.de), Robert Koch Institute, Germany

Sabrina Hense (HenseS@rki.de), Robert Koch Institute, Germany

Kari Kuulasmaa (kari.kuulasmaa@thl.fi), National Institute for Health and Welfare (THL), Finland

Ronan Lyons (r.a.lyons@swansea.ac.uk), Swansea University, UK

Madlen David (madlen.david@uba.de), German Environment Agency, Germany

Luigi Palmieri (luigi.palmieri@iss.it), Istituto Superiore di Sanià, Italy

Jean-Marie Robine (robinejm@gmail.com), Institut National de la Sante et de la Recherche Medicale (INSERM), France

Lau Caspar Thygesen (lct@si-folkesundhed.dk), National Institute of Public Health, University of Southern Denmark, Denmark

Thomas Ziese (ZieseT@rki.de), Robert Koch Institute, Germany

I. Introduction

Exchange of health information is needed to support policy making and international research on health related questions but also to allow benchmarking of national situations with neighbouring and other European countries. Depending on the use of health information, different levels of information are needed; individual level data, aggregated data on sub-groups or readily defined population level indicators.

II. Background

Health information can be obtained from different data sources; routine data collections (administrative data), electronic clinical data, disease specific registers, health surveys, epidemiological studies, and biobanks.

Routinely collected data (administrative data) are generated by healthcare systems as part of routine statistics including hospitalization information, information on medical operations, prescriptions of medications, mortality, etc. There are usually legal bases in countries for the collection of this type of data and the data are stored in centralised databases. These large databases typically include all individuals within a country who have been hospitalized, treated by a medical doctor etc.; are nationally representative and are regularly updated with new data.

In many settings there are also complete or partial electronic health record systems used to manage patients. Complete systems are less common and include the information collected by a variety of clinicians on patient history, symptoms, findings on examination, laboratory investigations, diagnoses, and treatment plans. Partial systems, such as the results of all laboratory and radiological investigations are much more common. Data from both systems are commonly used to feed into disease specific registers.

Disease specific registers such as for diabetes, cancer or rare diseases cover only one disease or disease group. These registers are usually established to monitor the incidence and prevalence or the natural course of the disease in question.

Survey data from different types of health surveys have specific information about health, diseases and determinants of health on an individual level. Health surveys can be health interview surveys, health examination survey, and environmental biomonitoring surveys or combinations of these. Surveys are conducted in a sample of the target population and are therefore more limited on coverage and size than administrative databases. However, when they are based on proper probability sampling and the response rate is high, the information they provide can be generalized to the entire underlying population. Surveys are usually conducted periodically on independent samples, but they can include also a longitudinal component.

Epidemiological studies are health studies on human populations concerning disease occurrence in specified population groups. Epidemiology considers the time and reason for the occurrence of specific health conditions. The results are used to prevent illness and to understand the reasons for the differences between population groups.

A biobank is a collection of biological samples such as blood, urine and other tissues, often complemented with related information such as socio-economic position, diagnosed diseases etc. Biological samples stored in biobanks can be used in biomedical research and retrospective laboratory analysis to determine new biomarkers. Many countries in Europe have biobanks. These biobanks can be specific for one study or hospital, or organization of joint biobanks for several instances. At the EU level, the European Research Infrastructure Consortium on Biobanking and BioMolecular Resources Infrastructure (BBMRI-ERIC) has been established to facilitate European level collaboration between biobanks. (Th. Mayrhofer 2016, Yuille 2007)

Health information obtained from these different data sources can be either raw data on individual level or data aggregated to some sub-groups such as sex, age groups or regions, or readily defined population-level indicators.

Individual level data comprise health information of a single patient or survey participant concerning his/her name, age, sex, diagnosis, medical history and other relevant information. If it is envisaged to record the course of the disease of a patient over time, it is necessary to collect individual data. This is also true if you want to communicate the results to each person. Ethical and legal issues of data collection are crucial when working with individual level data.

Aggregated data merge health information of multiple patients or survey participants and the collected information cannot be retraced to the individual data. Aggregated data are used in ecological studies and when analysing differences between countries or other population groups.

For research purposes, such as investigating relationships between exposures and onset, data on individual level is usually preferred. On the other hand, to support policy decisions, population sub-group specific indicators may be sufficient.

III. Anonymization of health information

Health data are always considered sensitive information and therefore safeguarding the privacy of individuals has an important role when handling this type of data. Data protection and ethical issues are considered in detail in the report from HA7 Ethical issues.

If individual level data are used and whenever data are transferred from one entity to the other, it is important to ensure the privacy of the individuals through anonymization. The term anonymous or anonymised data has been defined by the Working Partly on the protection of individuals with regard to the processing of personal data (Article 29) as *"any information relating to a natural person where the person cannot be identified, whether by the data controller or by any other persons, taking account of all the means likely reasonably to be used either by the controller or by any other person to identify that individual"*. In relation to anonymization, several concepts are used: de-identified, non-identifiable, irretrievable, unlinked, irreversible de-identification, unlinked-

anonymised, irreversibly anonymised and pseudonymised. (Elger et al 2010, Pfitzman 2006).

For use in this document, we define the following concepts:

1. *Personal data* allow identification of a natural person either directly or indirectly from the data. This does not only mean personal identification, that is name and address of the person, but also cases where sufficient other identifiers are present in the data which alone or in combination may lead to the identification of the individual. This can happen through merger of information provided by the individual on social media sites, e.g. inclusion of a date of birth, sex or small geographical area code.
2. *Pseudonymised data* have personal identifiers which only data controllers, with access to personal data, can link to a natural person.
3. *Reasonably anonymised data* means that no reasonable means of identification of specific individuals are available. This term is often used in relation to genetic data (WHO 2003).

There is a wide range of data anonymization techniques; substitution, shuffling, number variance, data variance, character masking, cryptographic techniques, public key techniques, message digest techniques, partial sensitivity and partial masking, masking based on external dependency, auxiliary anonymization techniques, alternative classification of data anonymization techniques and leveraging data anonymization techniques. (Raghunathan 2013)

IV. Data sharing models

Five models to share data may be distinguished.

1. *Open Data*. This involves sharing of data published on the internet and is usually restricted to strongly de-identified data, but is not always the case e.g. in social media. Open data sharing is a fairly common practice for genetic material unlinked to health records.

2. *Multi-site replication of analyses*. In this scenario individual level records do not cross organizational borders but analyses are replicated in multiple sites and aggregate statistics shared. E.g. calculation of standardized mortality ratios (SMRs) for specific conditions, or sharing of effect sizes and their standard errors from different cohorts.

3. *Transfer of strongly pseudonymised data*. These are data that have been stripped of strong personal identifiers, such as name, address, postcode (ZIP), date of birth and unique national or health service numbers. The data still contain unique records and therefore re-identification is a possibility, particularly if the data are subsequently linked to third party records.

4. *Secure analysis platforms*. Privacy protecting analysis platforms allow speudonymised data to be queried and prevent physical links to third party records. Analysis is usually restricted to accredited researchers (sometimes including health officials and Government statisticians). They generally only allow removal of summary non-disclosive statistics. Some platforms require the person to be in a specific physical location, sometimes monitored by CCTV, whereas others can be remotely accessed.

5. *Federated analysis*. Federated analysis of data held in multiple locations is possible using technologies, such as DataShield (http://cran.datashield.org/). DataSHIELD is an R library that enables the remote and non-disclosive analysis of sensitive research data. Users are not required to have prior knowledge of R programme (https://www.r-project.org/). Data need to be completely harmonized in advance of analysis.

Examples of the five models in operation.

A. Open Data: e.g. MINE website (http://www.exploredata.net/Downloads/Gene-Expression-Data-Set) or DisGenet (http://www.disgenet.org/web/DisGeNET/menu;jsessionid=hcdh9bt2r6rg1rs6q5msmstpq)

B. Multi-site replication of data: see example of analysis of cardiovascular disease from the 2015 UK Farr Institute Annual report:

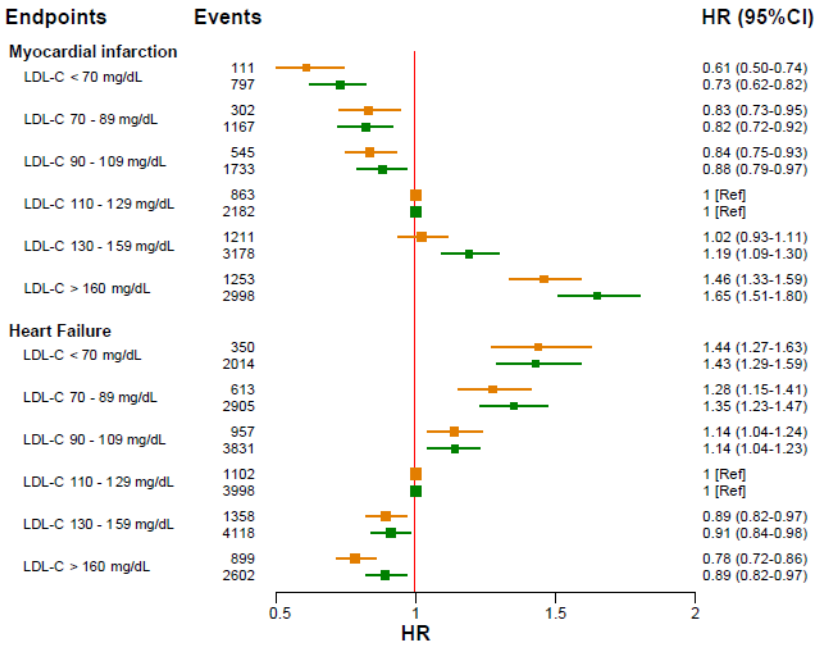

Figure: Relationship between LDL-Cholesterol levels and myocardial infarction and heart failure in independent analyses of GP data (500,000 CPRD records in England and 700,000 SAIL records in Wales).

C. Transfer of de-identified data: for example, the mortality and hospitalization data used by EU (http://ec.europa.eu/health/data_collection/portal/index_en.htm), OECD (https://data.oecd.org/health.htm) and WHO (http://www.exploredata.net/Downloads/WHO-Data-Set).

D. Privacy protecting Platforms: for example, the Secure Anonymised Information Linkage (SAIL) system (www.saildatabank.com) Lyons et al (2014); Administrative Data Research Network (https://adrn.ac.uk/); and Dementias Research Platform (http://www.dementiasplatform.uk/) in the UK.

E. Federated analysis: DataSHIELD is currently used by the Healthy Obese Project and the Environmental Core Project (BioSHaRE-EU) for the federated analysis of ten data sets across eight European countries (https://www.bioshare.eu/).

V. Aim

The aim of this report is to identify on which level health information has been collected and shared; individual, aggregated or indicators at the population-level. Also the methods used for sharing the health information between organizations/stakeholders etc. are evaluated.

VI. Approach

A questionnaire relating to the horizontal activities of the BRIDGE Health Project (http://www.bridge-health.eu) was mailed to 23 persons in November 2015. The questionnaire included following questions about health data exchange methods used:

1. In your area/project, have you shared/exchanged collected health information between organizations and/or 3rd parties?
    a. If yes, have you developed a data (information) sharing rules?
2. Are you aware of any legislation/regulations (national or international) which might prevent or made exchange of health information difficult?

Replies were obtained from 19 persons.

Information was also obtained from project web sites when available and through personal contacts.

VII. Results

A. Health information and level of anonymization

Table 1. lists what kind of health information, and on which level of aggregation and anonymization has been shared between Partners in previous or ongoing EU funded projects on health information. In most studies, pseudonymisation of data has been shared, especially when individual level data have been collected.

Table 1. Level of aggregation and anonymization of data shared between partners in EU funded projects

Project	Data source	Aggregation level	Level of anonymization	Type of data	Data management structure
EHES	Survey	Individual level	Pseudonymised	Health and health determinants, socio-economic background information	Centralized database (relational database at THL)
COPHES/ DEMOCOPHES	Survey	Individual level	Pseudonymised	Environmental health (nutrition, smoking behaviour, exposure-relevant behaviour, occupation), socio-economic background information, and human biomonitoring information	
Euro-Peristat		Aggregated to different levels	Anonymised	Health information related to perinatal health	
JA-ECHIM	Multiple, depending on indicator: e.g. Eurostat, WHO, OECD, international databases	Indicators	Anonymised	Multiple, depending on indicators: e.g. HIS or other surveys, administrative data, population statistics, registers, specific databases	Aggregated data and metadata are available online from the ECHI Database (http://ec.europa.eu/health/indicators/indicators/index_en.htm)
EuroSafe/IDB	Administrative	Individu	Pseudonymis		Centralized database

Project	Data source	Aggregation level	Level of anonymization	Type of data	Data management structure
(Injury Data Base)	data	al level	ed		(https://ec.europa.eu/health/data_collection/databases/idb_en)
EUROHOPE	Administrative data	Individual level	Pseudonymized		
ECHO	Administrative data	Individual level	Pseudonymized in origin – anonymized later on the ECHO data model	Health status and health system performance, health care service performance, variations in health care services	Centralized database (relational data model at IACS)

B. Data exchanges

Table 2. provides a summary about used data sharing rules and methods among previous EU projects. In most projects, only aggregated level data or readily defined indicators are shared between organizations. In few studies individual level data have also been shared.

Each time when individual level data are shared, a special attention has been paid to requirements set by national legislation. Written data transfer agreements have been prepared between data owners and the organization managing the centralized database. Only in the EHES Project, a protocol for providing pseudonymized individual level data from centralized database to research groups (3rd parties) for further analysis has been developed and adapted. In other projects working with individual level data, only aggregated data have been shared with 3rd parties.

Table 2. Level of shared data and data sharing methods

Project	Data ownership	Level of shared data	Written data transfer agreement			Data use request (included components)
			From country to centralized database	From centralized database to research group	From country to research group	
EHES	Stays within data provider	Individual level	Yes	Yes, if data are transferred to a different organization.	Yes, if data are transferred to a different organization.	Written proposal including purpose of the analysis, place for the analysis, tentative manuscript group, timeline for the work. Proposal approved by Publication Committee and each relevant data owner.
COPHES/ DEMOCOPHES	Stays within data provider	Individual level				
Euro-Peristat		Indicators	No	No	No	No
JA-ECHIM		Indicators	No	No	No	No
EuroSafe/ IDB	Stays within data provider	Individual level	Yes	No	No	Secure access to centralized database is made available to all the national data administrators.
		Aggregated data		Yes		
EHLEIS		Indicators	No	No	No	No
EUROCISS		Aggregated register data				
EUROHOPE	Stays within data provider	Individual level	Yes			
ECHO	Stays within data provider	Individual level?	No	No	Yes (It depends on	Case-to-case agreement with each partner or institution responsible for

					the country transferring the data)	the data on each country. Written proposal of analysis, place for the analysis tentative manuscript group, time-lines for the work and system level security policy (ARiHSP group – IACS)

VIII. EU data protection regulation

Under EU law, personal data can only be gathered legally under strict conditions and for a legitimate purpose. Persons and organizations that collect and manage personal information are under an obligation to protect it from misuse and protect the rights of the data owners.

The General Data Protection Regulation (GDPR) strengthens data protection for individuals within the EU and addresses the export of data outside the EU. It was adopted on 27 April 2016 and will take effect on 25 May 2018, after a two-year transition by Member States and without requiring any enabling legislation to be passed by governments. The regulation applies if the data controller or processor of personal data is based in the EU or outside, with the exception of national security or law enforcement activities. Personal data are defined as any information relating to an individual, whether it relates to his or her private, professional or public life and can be anything from a name, photo, email address, bank details, material on social network sites, medical information or a computer's IP address. The Directive requires privacy by design and default.

The regulation is available at the Official Journal of the European Union L119/1 4/5/2016 (http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN).

The main points relevant to BRIDGE Health are summarized below with the relevant paragraphs included in the Appendix.

Consent is generally required for the processing of personal data. Pseudonymisation is strongly encouraged as is the removal of sufficient identifying information to make the information truly anonymous and hence not subject to the GDPR. A combination of pseudonymisation and the use of privacy protecting restricted access platforms can achieve this aim (Jones, 2014). There are specific derogations from the requirement for consent, where this is not feasible, for processing personal data for medical research and statistical purposes.

The practical implications of the GDPR and increasing consideration of data protection and ethical issues around the management of data by many organisations and member states is likely to influence the extent to which data, whether identifiable, pseudonymised or anonymized are transferred across borders.

IX. Future developments

Already now, but more so in the near future, terms such as 'MyData', 'Big data' and 'Open data' will find their way also to 'Health Information'. These will set new challenges for the protection and management of personal data and for the acceptability and legality of data exchange.

A. MyData

MyData is a generally used term which refers to practices on how personal data are managed and processed. Personal data may relate to one's health, use of public services, education, retail experience (what he/she has bought from shops), use of media, web services, self-measurement such of data from activity trackers, etc. On a daily basis, huge amounts of information about one's activities are collected by different databases, particularly by social media and retail organizations, but only in limited cases does the person have full control over the uses and distribution of the collected data. User terms and conditions are so long that very few people read them and sign up to unconstrained uses of their data in the commercial sector. The ideology behind MyData is that all collected information could be used more effectively by linking them together at the individual level and providing individuals better with control over their own information. (Poikola et al. 2015, https://www.midata.coop/index.html)

In theory, this provides many opportunities for the generation of health information but at the same time raises many questions relating to data protection and ethical use of data. (Hafen et al 2014). MyData ideology is closely linked to 'Big Data' and 'Open Data' concepts and discussions.

B. Big Data

'Big Data' refers to datasets which are large not only in volume but rather in complexity.

There are many EU funded projects dealing with this issue, including BIG – the Big Data public private forum (http://www.big-project.eu/) which ran from 2012-14. BIG worked towards the definition and implementation of a clear strategy that tackles the necessary efforts in terms of research and innovation, but it also aims at providing a major boost for technology adoption and supporting actions from the European Commission in the successful implementation of the *Big Data* economy. A number of BIG publications from 2013 and 201 are available at (http://www.big-project.eu/publications).

Uses of 'Big Data' can already be seen in health informatics research, particularly in genetics (Hawking et al 2010) and register based health research.

A number of jurisdictions and regions now bring together large scale complex multi-modal data (Big Data) from many sources, including electronic health records, disease registers, population surveys and disease registers at the individual level, whilst still protecting privacy, in order to be able to support the 21st century research paradigms focused around discovery science, precision medicine, multi-organisational public health interventions, learning health systems and citizen driven health. Single disease or single data source systems rarely have the capabilities required to answer complex policy relevant questions in sufficient depth. Examples of such systems are to be found in an increasing number of countries, for example, Finland (Gissler 2014), Sweden (Cnudde 2016), Denmark (Schmidt 2015), Norway (Moller 1994), United Kingdom (Muckerjee 2016; Hutchings 2016), Australia (Holman 2008), Canada (Chen 2017), New Zealand (Chan 2014) and the US (Brownstein 2010).

C. Open Data

Open data are data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike (Open Data Handbook) .

There are many benefits from Open Data including, low cost and universal availability that stimulate the development of services to produce efficient, analytical tools and services. Key components of efficiency in 'Open Data' are the adoption of common standards and formats, which are essential to interoperability, which in turn supports reuse of data. Open data formats are commonly used in many settings, including formats for collection and transfer of medical image data.

Open data requires consideration of privacy protection and hence generally only completely anonymised individual level information or non-personal information are shared through open data platforms, unless where individuals have given informed consent, such as in a number of cohort studies. However, whenever data are made open or in the 'wild', then there is potential for deliberate or inadvertent re-identification through linkage to information in third party sites. There is a growing industry of re-identification scientists who commonly work for two groups, national security agencies and commercial e-commerce organisations. Both attempt to piece together data items on individuals to identify security risks or to more precisely target marketing. Those who use well known internet search engines or social media sites will have a grasp on the efficiency of their identity algorithms. The most high profile breaches of privacy protection to date all relate to linking presumed pseudonymised open data from more than two sites, enabled by the content of social media sites.

X. Implications and limitations

Europe has many policy and research needs that can only be addressed by integrating data and expertise from many sources. It needs to do this by implementing systems that are not only legal but also have the support of communities and the medical and regulatory experts who often decide whether or not to permit the reuse of data for these purposes. Current systems are helpful but are not aligned to make maximum use of available data and expertise. There is a tension between the super-national centralisation and efficient use of data and the risks to privacy protection and reduction in population acceptance of this approach. Given that regulations and public opinion are not very favourable to super-national transfer of data in circumstances where their health is not immediately threatened (communicable disease is an exception) research groups have started to adopt federated approaches to data sharing with increased harmonisation of data and analytical algorithms at source. The best ways of achieving such harmonisation and collaboration from technical, efficiency and acceptability perspectives is still somewhat uncertain and points to the need for a future European Research Consortium to develop, implement and evaluate such solutions.

XI. Conclusions

For efficient use of health information both for research and support of public health policy decisions, health data sharing/exchange is essential. For European Health Information System, needs for health data by different users should to be identified including do they need individual level data or is aggregated level information sufficient for them.

It may well be, that Open Access health information platform with aggregated level information is sufficient for some user groups. Then the question becomes how to obtain this aggregated level information in standardized way and solutions for this has to be sorted out.

Obviously, there will also be user groups, especially researches who would need the access to individual level data. This will require establishment of data sharing system and rules which ensures data protection and data ownership and same time makes use of efficient data sharing.

For health information sharing/exchange, there is no one solution fits all situations structure and it may be that different solutions are needed for different data sources and data provides.

XII. References

- Allen J, Holman CD, Meslin EM, Stanley F (2013). Privacy protectionism and health information: is there any redress for harms to health? J Law Med; 21(2):473-85.
- Article 29 Working Party (2007), Opinion No 4/2007 on the concept of personal data, WP 136. Available from: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf
- Brownstein JS, Murphy SN, Goldfine AB, Grant RW, Soldo M, Gainer V, Colecchi JA, Dubey A, Nathan DM, Glaser JP, Kohane IS (2010). Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. Diabetes Care; 33:526-31
- Chan WC, Jackson G, Wright CS, Orr-Walker B, Drury PL, Boswell DR, Ai Wei Lee M, Papa D, Jackson R. (2014) The future of population registers: linking routine health datasets to assess a population's current glycaemic status for quality improvement. BMJ Open; 4:e003975. doi:10.1136/bmjopen-2013- 003975
- Chen H, Kwong JC, Copes R, Tu K, Villeneuve PJ, van Donkelaar A, Hystad P, Martin RV, Murray BJ, Jessiman B, Wilton AS, Kopp A, Burnett RT (2017). Living near major roads and the incidence of dementia, Parkinson's disease, and multiple sclerosis: a population-based cohort study. Lancet; Jan 4. doi:10.1016/S0140-6736(16)32399-6
- Cnudde P, Rolfson O, Nemes S, Karrholm J, Rehnberg C, Rogmark C, Timperley J, Garellick G (2016). Linking Swedish health data registers to establish a research

database and a shared decision-making tool in hip replacement. BMC Musculoskelet Disord.; 17:414. doi:10.1186/s12891-016-1262-x

- Elger BS, Iavindrasan J, Lo Iocono L et al. (2010) Strategies for health data exchange for secondary, cross-institutional clinical research. Computer methods and programs in biomedicine, 99: 230-251
- Gissler M, Klemetti R, Sevon T, Hemminki E (2004). Monitoring of IVF birth outcomes in Finland: a data quality study. BMC Medical Informatics and Decision Making; 4:3. doi:10.1186/1472-6947-4-3
- Hafen E, Kossmann D, Brand A. (2014) Health Data Cooperatives – Citizen Empowerment. Methods Inf Med; 53:82-86, doi:10.3414/ME13-02-0051
- Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nature; 11:476-486
- Holman CDAJ, Bass JA, Rosman DL, Smith MB, Semmens JB, Glasson EJ, Brook EL, Trutwein B, Rouse IL, Watson CR, de Klerk NH, Stanley FJ (20008). A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Australian Health Review; 32(4):766–777. doi:10.1071/AH080766
- Hutchings H, Evans A, Barnes P, Demmler JC, Heaven M, Healy MA, James-Ellison M, Lyons RA, Maddocks A, Paranjothy S, Rodgers SE, Dunstan F (2016). Residential Moving and Preventable Hospitalizations. Pediatrics. doi:10.1542/peds.2015-2836
- Jones KH, Ford DV, Jones C, D'Silva R, Thompson S, Brooks CJ, Heaven ML, McNerney CL, Lyons RA (2014). The Secure Anonymous Information Linkage (SAIL) Gateway: a case study describing a remote access system for health-related research and evaluation. Journal of Biomedical Informatics 01/2014; DOI:10.1016/j.jbi.2014.01.003.
- Lyons RA, Ford DV, Moore L, Rodgers SE (2014). Using data linkage to measure the population health impact of non-healthcare interventions. The Lancet; 383:1517-1518. http://dx.doi.org/10.1016/S0140-6736(13)61750-X
- Moller H, Mellemgaard A, Lindvig K, Olsen JH (1994). Obesity and cancer risk: a Danish record-linkage study. European J Cancer; 30(3):344-50. doi: 10.1016/0959-8049(94)90254-2
- Mukherjee M, Stoddart A, Gupta RP, Nwaru BI, Farr A, Heaven M, Fitzsimmons D, Bandyopadhyay A, Aftab C, Simpson C, Lyons RA, Fischbacher C, Dibben C, Shields M, Phillips C, Strachan D, Davies G, McKinstry B, and Sheikh A (2016). The epidemiology, healthcare and societal burden and costs of asthma in the UK and its member nations: analyses of standalone and linked national databases. BMC Medicine; 14:113. doi:10.1186/s12916-016-0657-8
- Ohm P (2010). Broken promises of privacy: responding to the surprising failure of anonymization. Available at: http://www.uclalawreview.org/pdf/57-6-3.pdf
- Open Data Handbook. Available at: http://opendatahandbook.org/guide/en/

- Pfitzmann A, Hanse M. (2006) Anonymity, unlinkability, unobservability, pseudonymity, and identity management – a consolidated proposal for terminology. http://dud.inf.tu-dresden.de/Anon_Terminology.shtml
- Poikola A, Kuikkaniemi K, Honko H. (2014) MyData – A Nordic Model for human-centered personal data management and processing. Ministry of Transport and Communications. Available at: https://www.lvm.fi/documents/20181/859937/MyData-nordic-model/2e9b4eb0-68d7-463b-9460-821493449a63?version=1.0
- Raghunathan B (2013). The complete book of data anonymization. From planning to implementation. CRC Press. ISBN 978-1-4398-7730-2
- Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT (2015). The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol.; 7:449–490. doi: 10.2147/CLEP.S91125
- Th. Mayrhofer M, Holub P, Wutte A, Litton J-E (2016). BBMRI-ERIC: the novel gateway to biobanks. Bundesgesundheisbl; 59:379-384, doi:10.107/s00103-015-2301-8
- WHO (2013). The World Health Organization (European Partnership on Patientes' Rights and Citizens' Empowerment), Genetic Databases, Assessing and the Impact on Human Rights and Patient Rights, World Health Organization, Geneva.
- Yuille M, van Ommen G-J, Bréchot C et al. (2007) Biobanking for Europe. Briefings in bioinformation; 9(1):14-24, doi:10.1093/bib/bbm050
- Weber GM, Mandl KD, Kohane IS (2014). Finding the missing link for big biomedical data. JAMA; 311(24):2479-80.

XIII. Appendix. Particularly relevant sections of the EU General Data Protection Regulation

Paragraph numbers are included in parentheses

(23) The principles of data protection should apply to any information concerning an identified or identifiable natural person. Data which has undergone pseudonymisation, which could be attributed to a natural person by the use of additional information, should be considered as information on an identifiable natural person. To determine whether a person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by any other person to identify the individual directly or indirectly. To ascertain whether means are reasonable likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the processing and technological development. The principles of data protection should therefore not apply to anonymous information, that is information which does not relate to an identified or identifiable natural person or to data rendered anonymous in such a way that the data subject is not or no longer identifiable. This Regulation does therefore not concern the processing of such anonymous information, including for statistical and research purposes.

(23a) The application of pseudonymisation to personal data can reduce the risks for the data subjects concerned and help controllers and processors meet their data protection obligations. The explicit introduction of 'pseudonymisation' through the articles of this Regulation is thus not intended to preclude any other measures of data protection.

(23c) In order to create incentives for applying pseudonymisation when processing personal data, measures of pseudonymisation whilst allowing general analysis should be possible within the same controller when the controller has taken technical and organisational measures necessary to ensure, for the respective processing, that the provisions of this Regulation are implemented, and ensuring that additional information for attributing the personal data to a specific data subject is kept separately. The controller processing the data shall also refer to authorised persons within the same controller.

(27) Personal data concerning health should include all data pertaining to the health status of a data subject which reveal information relating to the past, current or future physical or mental health status of the data subject; including information about the individual collected in the course of the registration for and the provision of health care services as referred to in Directive 2011/24/EU to the individual; a number, symbol or particular assigned to an individual to uniquely identify the individual for health purposes; information derived from the testing or examination of a body part or bodily substance, including genetic data and biological samples; or any information on e.g. a disease, disability, disease risk, medical history, clinical treatment, or the actual physiological or biomedical state of the data subject independent of its source, such as e.g. from a physician or other health professional, a hospital, a medical device, or an in vitro diagnostic test.

(31) In order for processing to be lawful, personal data should be processed on the basis of the consent of the person concerned or some other legitimate basis, laid down by law, either in this Regulation or in other Union or Member State law as referred to in this Regulation, including the necessity for compliance with the legal obligation to which the controller is subject or the necessity for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract.

(42) Derogating from the prohibition on processing sensitive categories of data should also be allowed when provided for in Union or Member State law and subject to suitable safeguards, so as to protect personal data and other fundamental rights, where grounds of public interest so justify, in particular processing data in the field of of employment law, social protection law including pensions and for health security, monitoring and alert purposes, the prevention or control of communicable diseases and other serious threats to health. This may be done for health purposes, including public health and the management of health-care services, especially in order to ensure the quality and cost-effectiveness of the procedures used for settling claims for benefits and services in the health insurance system, or for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes. A derogation should also allow processing of such data where necessary for the establishment, exercise or defence of legal claims, regardless of whether in a judicial procedure or whether in an administrative or any out-of-court procedure

(42a) Special categories of personal data which deserve higher protection, may only be processed for health-related purposes where necessary to achieve those purposes for the benefit of individuals and society as a whole, in particular in the context of the management of health or social care services and systems including the processing by the management and central national health authorities of such data for the purpose of quality control, management information and the general national and local supervision of the health or social care system, and ensuring continuity of health or social care and cross-border healthcare or health security, monitoring and alert purposes, or for archiving purposes in the public interest or scientific and historical research purposes or statistical purposes based on Union or Member State law which has to meet an objective of public interest, as well as for studies conducted in the public interest in the area of public health. Therefore this Regulation should provide for harmonised conditions for the processing of special categories of personal data concerning health, in respect of specific needs, in particular where the processing of these data is carried out for certain health-related purposes by persons subject to a legal obligation of professional secrecy. Union or Member State law should provide for specific and suitable measures so as to protect the fundamental rights and the personal data of individuals. Member States should be allowed to maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or health data. However, this should not hamper the free flow of data within the Union when those conditions apply to cross-border processing of such data.

(42c) The processing of special categories of personal data may be necessary for reasons of public interest in the areas of public health without consent of the data subject. This processing is subject to suitable and specific measures so as to protect the rights and

freedoms of individuals. In that context, ‘public health’ should be interpreted as defined in Regulation (EC) No 1338/2008 of the European Parliament and of the Council of 16 December 2008 on Community statistics on public health and health and safety at work, meaning all elements related to health, namely health status, including morbidity and disability, the determinants having an effect on that health status, health care needs, resources allocated to health care, the provision of, and universal access to, health care as well as health care expenditure and financing, and the causes of mortality. Such processing of personal data concerning health for reasons of public interest should not result in personal data being processed for other purposes by third parties such as employers, insurance and banking companies.

(61) The protection of the rights and freedoms of individuals with regard to the processing of personal data require that appropriate technical and organisational measures are taken to ensure that the requirements of this Regulation are met. In order to be able to demonstrate compliance with this Regulation, the controller should adopt internal policies and implement measures, which meet in particular the principles of data protection by design and data protection by default. Such measures could consist inter alia of minimising the processing of personal data, pseudonymising personal data as soon as possible, transparency with regard to the functions and processing of personal data, enabling the data subject to monitor the data processing, enabling the controller to create and improve security features. When developing, designing, selecting and using applications, services and products that are either based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations.The principles of data protection by design and by default should also be taken into consideration in the context of public tenders.

(125) The processing of personal data for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes should be subject to appropriate safeguards for the rights and freedoms of the data subject pursuant to this Regulation. These safeguards should ensure that technical and organisational measures are in place in order to ensure, in particular, the principle of data minimisation. The further processing of personal data for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes is to be carried out when the controller has assessed the feasibility to fulfill those purposes by processing data which does not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist (such as, for instance, pseudonymisation of the data). Member States should provide for appropriate safeguard to the processing of personal data for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes. Member States should be authorised to provide, under specific conditions and in the presence of appropriate safeguards for data subjects, specifications and derogations to the information requirements, rectification, erasure, to be forgotten, restriction of processing and on the right to data portability and the right to object when processing personal data for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes. The conditions and safeguards in question may entail specific procedures for data subjects to exercise those rights if this is appropriate in

the light of the purposes sought by the specific processing along with technical and organisational measures aimed at minimising the processing of personal data in pursuance of the proportionality and necessity principles. The processing of personal data for scientific purposes should also comply with respect to other relevant legislation such as on clinical trials.

(125aa) By coupling information from registries, researchers can obtain new knowledge of great value when it comes to e.g. widespread diseases as cardiovascular disease, cancer, depression etc. On the basis of registries, research results can be enhanced, as they draw on a larger population. Within social science, research on the basis of registries enables researchers to obtain essential knowledge about long-term impact of a number of social conditions e.g. unemployment, education, and the coupling of this information to other life conditions. Research results obtained on the basis of registries provide solid, high quality knowledge, which can provide the basis for the formulation and implementation of knowledge-based policy, improve the quality of life for a number of people, and improve the efficiency of social services etc. In order to facilitate scientific research, personal data can be processed for scientific research purposes subject to appropriate conditions and safeguards set out in Member State or Union law.

(126c) Where personal data are processed for statistical purposes, this Regulation should apply to that processing. Union law or Member State law should, within the limits of this Regulation, determine statistical content, control of access, specifications for the processing of personal data for statistical purposes and appropriate measures to safeguard the rights and freedoms of the data subject and for guaranteeing statistical confidentiality. Statistical purposes mean any operation of collection and processing of personal data necessary for statistical surveys or for the production of statistical results. These statistical results may further be used for different purposes, including a scientific research purpose. Statistical purposes mean any operation of collection and processing of personal data necessary for statistical surveys or for the production of statistical results. The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the data are not used in support of measures or decisions regarding any particular individual.