

Health Data for Public Health: Towards New Ways of Combining Data Sources to Support Research Efforts in Europe

A. Burgun¹, E. Bernal-Delgado², W. Kuchinke³, T. van Staa⁴, J. Cunningham⁴, E. Lettieri⁵, C. Mazzali⁵, D. Oksen⁶, F. Estupiñan², A. Barone⁷, G. Chêne⁸

¹ Inserm, UMR 1138, Centre de Recherche des Cordeliers, AP-HP, Paris Descartes University, Paris, France

² Institute for Health Sciences in Aragon (IACS), BridgeHealth Consortium, Zaragoza, Spain

³ University of Dusseldorf, Dusseldorf, Germany

⁴ Health eResearch Centre, Farr Institute, University of Manchester, Manchester, United Kingdom

⁵ Politecnico Milano, Milano, Italy

⁶ Public Health Institute, Inserm, AVIESAN, Paris, France

⁷ Lombardia Informatica, Milano, Italy

⁸ Inserm, UMR 1219, CIC1401-EC, Univ. Bordeaux, ISPED, CHU Bordeaux, Bordeaux, France

Summary

Objectives: To present the European landscape regarding the re-use of health administrative data for research.

Methods: We present some collaborative projects and solutions that have been developed by Nordic countries, Italy, Spain, France, Germany, and the UK, to facilitate access to their health data for research purposes.

Results: Research in public health is transitioning from siloed systems to more accessible and re-usable data resources. Following the example of the Nordic countries, several European countries aim at facilitating the re-use of their health administrative databases for research purposes. However, the ecosystem is still a complex patchwork, with different rules, policies, and processes for data provision.

Conclusion: The challenges are such that with the abundance of health administrative data, only a European, overarching public health research infrastructure, is able to efficiently facilitate access to this data and accelerate research based on these highly valuable resources.

Keywords

Data sources; public health informatics; Europe; health administrative data; secondary use of health data

Yearb Med Inform 2017;235-40

<http://dx.doi.org/10.15265/Y-2017-034>

Published online August 18, 2017

1 Introduction

Research in Public Health basically needs data of high quality. It is currently transitioning from multiple, stand-alone, siloed systems to interconnected resources and systems. Such a transition involves major challenges for public health informatics, especially in Europe.

- Data sharing is crucial and already strongly endorsed by the H8 group of global health organisations. “*Informatics and the ability to mine large datasets and combine them with information from many other sources present a huge potential to advance developments in public health.*”[1]. Several research communities, as in the HIV field, already compile and harmonize their data [2]. This process should be generalised to all data useful for public health research, given the expected benefits of data sharing [3].
- Data sharing of routinely generated health data requires far more efforts than simply uploading a dataset to a common repository. Three major obstacles need to be overcome [4]: (i) fragmentation of data, (ii) respect of the confidential nature of health data, (iii) insufficient workforce to mine health data. A first crucial step is

needed at the national level with record linkage and integration of all repositories as needed to provide comprehensive views of health information.

- Health systems typically have separate information pipelines (i) for monitoring public health and policy making and (ii) for research. These pipelines share common data sources, but they usually involve different groups and duplicate data extraction and analytics.

In this paper, we report on several European initiatives in public health informatics and the re-use of data for research.

2 Re-use of Health Data in Europe for Public Health

2.1 The Potential of Health Administrative Data

Several EU projects have promoted the re-use of existing medical databases for research purposes, e.g. [5, 6]. Health administrative data exhibit major strengths for research, especially when:

- The entire population of a region or a

country is covered by the data collection.

- Construction of longitudinal histories of individuals is possible.
- Multiple years of data allow powerful identification of trends over time.

Data are accessible with low research costs. Data can therefore be used to increase scientific knowledge in a number of areas: epidemiological surveillance, health determinants, health services research, economic evaluations, risk/benefit ratio of drugs and medical devices in “real” population, comparative effectiveness research, and prevention. The re-use of such data allows especially the adequate powered research on rare diseases and rare events.

2.2 European Landscape on the Re-use of Health Administrative Data

The BridgeHealth consortium (<http://www.bridge-health.eu>) has established a list of European projects able to yield indicators valuable for stakeholders to make decisions in public health, like Euro-Peristat (<http://www.europeristat.com/>) in perinatal health [7]. Other projects like ECHO [8] and Euro-HOPE [9] focus on healthcare performance comparisons. More focused information systems and registries were created for specific conditions like injuries (http://ec.europa.eu/health/data_collection/databases/idb/), and diabetes (<http://www.eubiroid.eu/>), or specific types of data, such as Diagnostic Related Groups (<http://eurodrg.eu/>). Finally, projects such as SHARE, EHLEIS or INEQ-CITIES focus on the social determinants of health by collecting data from several cities [10].

The heterogeneity of information across Europe has been a major concern to conduct cross-national studies. Projects like EU-ADR [11], EUROmedICAT [12], PROTECT [13], and EMIF [6] have developed their own solutions, mainly based on distributed networks and common data models like OMOP [14]. Thus, several projects in Europe allow cross-national information. However, there is currently no pan-European solution or comprehensive project addressing the access to health administrative data for research.

3 Accessing Health Data for Research in Europe

This section provides an overview of the infrastructures put in place by some European countries and the different models implemented to provide access to health administrative data for research purposes. In several countries, the infrastructures are not limited to health administrative data (strictly speaking) but rather extended to other data generated in routine care (including patient records), or produced for public health (like registries). The publications and projects described in this paper exemplify the potential of such works for research in Public Health, but are not based on a systematic review.

3.1 Nordic Countries

Nordic countries, i.e., Denmark, Sweden, Norway, Finland, and Iceland, have a common history with similar political systems and a national level taxpayer financed health insurance coverage. In Sweden, Norway, and Denmark, national health registries are administered by a single authority and contain information covering the entire population regarding diverse aspects of population health: cancer, in and out patient information, drug prescription, causes of death, and biobank information. Administrative data on migration and socioeconomic status are also collected at a national level, with data available since the 19th century and follow-up of individuals from birth or immigration, to death or emigration [15,16]. National health registers are available for public research. Private entities can also access data through public research institutions for research purposes. Denmark is currently ahead in the digitized access to data. Danish public researchers can indeed access all health registers linked together via an online tool called ForskerMaskinen or the ResearchEngine. Researchers can thus access the entire set of health registers of the Danish population online, although exports are limited to the sole results and aggregated data. Moreover, the joint Nordic project Tryggve is developing solutions to share data across national borders (<https://neic.nordforsk.org/2016/12/13/tryggve-takes-care-of-sensitive-data.html>).

An advantage of the Nordic registers in terms of accessibility is that different data sources can be linked via personal identification numbers that each individual is assigned to by birth or immigration. Dataset for analysis can therefore include very comprehensive data on a patient trajectory within the health system. As an example, in one study conducted within the Euro-Peristat project, three quarters of the evidence found were derived from Nordic countries data [7]. Researchers in Nordic countries can access data to set up nationwide patient cohort studies, adequately powered in terms of sample size and follow-up, from birth to death, in order to study the role of a large panel of health determinants [17, 18].

3.2 The United Kingdom

The *care.data* programme was set up to create a central database covering the overall English population (<https://www.england.nhs.uk/ourwork/tsd/care-data/>). However, it was cancelled after substantial concerns expressed by the public and general practitioners [19, 20]. The Clinical Practice Research Datalink (CPRD) is an initiative related to the *care.data* programme. It evolved from the General Practice Research Database (GPRD) to collate data from a large number of general practices in England and link this data to hospital data, around 50 disease registries and clinical audits, UK Biobank with genetic information and the loyalty cards of a large supermarket chain. CPRD currently includes data from over 600 general practices and over 10 million patients. It is very widely used internationally for epidemiological research [21], and even for clinical trials. This includes the recruitment of patients through their electronic health records (EHRs) for a pharmacogenetic study of statin-induced myopathy [22], the conduct of cluster randomised trials [23], and pragmatic point-of-care trials randomising patients between different standard treatments [24].

Other initiatives in the UK include (i) the Welsh Secure Anonymous Information Linkage System (SAIL) with a platform that operates a remote access system providing a secure data access to approved

users and data analysis tools [25], (ii) the Scottish Health Informatics Programme (SHIP)[26], and (iii) the Connected Health Cities (CHC) programme (<https://www.connectedhealthcities.org/>) in the North of England. In SAIL, SHIP and CHC, there are substantial public engagement programmes aiming at better understanding the publics' preferences and concerns related to the sharing of health data. Transparency is indeed crucial about information security, dynamic consent with the ability to opt-out of some specific uses of data, scientific transparency and reliability, and systematic public engagement [27].

3.3 France

There are more than 900 available databases for research in France (<https://epidemiologie-france.aviesan.fr/en/epidemiology/home>). Among them, CepiDC is a national database that gathers causes of death, and the National Health Insurance Inter-Scheme Information System (SNIIRAM) is a database managed by the French National Health Insurance Fund (CNAMTS) holding information on more than 65 million individuals, including all reimbursement data for medical consultations, drugs and procedures in ambulatory medicine, and hospital discharge summaries. Researchers can access the data after being accredited through a specific process. The number of publications based on this data has increased in the last years, covering topics in epidemiology and pharmacoepidemiology [28], health care costs and health care assessment [29]. Several publications were issued from the CNAMTS itself, estimating the medical and economic burden of chronic diseases, and analysing social determinants of health [30], and comparing SNIIRAM to dedicated registries [31].

However, several obstacles still limit the access to data. An internal survey conducted by INSERM in 2015 (http://www.europeristat.com/images/Paris_2016_meeting/Collaborations/burgun-5apr16.pdf) showed that an average of 18 months and no less than 16 formal steps were needed to access such data, and approvals for record linkage were even more difficult to

obtain. This system has moved in 2016 to a more inclusive data repository, with claims data, hospital coded data sets, and causes of death, called the National Health Data System (SNDS, for “Système National des Données de Santé”). INSERM is setting up an infrastructure to facilitate the access to the SNDS for research while preserving a high level of security and privacy.

3.4 Italy

Each Italian Region routinely collects information to be shared with the central government on healthcare activities. The use of regional data for research purposes is usually granted within specific projects. It is supported by different IT solutions, e.g., the Lombardy Region is studying a solution similar to the Danish *Forskermaskine*. Data collected through the New National IT System (NSIS) is used to evaluate the quality and appropriateness of healthcare services. Several informative flows centred on both the provider or the citizen about hospitalizations, ambulatory care, and socio-sanitary activities are stored in the NSIS. Moreover, the Italian National Institute of Statistics (ISTAT) maintains longitudinal databases for monitoring and evaluating the healthcare system and citizens' health conditions. While the NSIS is mainly based on health administrative data, other databases are used for research purposes, among which those maintained by Health Search, the research institute of the Italian Association of General Practitioners, with more than 100 publications (<https://healthsearch.it/>), and Pedianet, a project promoted by an independent network of Italian paediatricians (<http://www.pedianet.it/en/>), with also a great number of publications, e.g., [32]. Administrative data, collected at the regional level, are at the basis of an important evaluation program at the national level, the National Outcome Project (PNE), which includes the evaluation of healthcare processes, interventions, providers, and access to healthcare services. Moreover, those data are at the basis of several research projects and publications [33], mostly in pharmacoepidemiology, health care management, and epidemiology, including studies on diseases with very low incidence [34].

3.5 Spain

The Spanish Ministry of Health (MSSSI) in coordination with the 17 Regional Health Departments has the responsibility of the collection of health data and the maintenance of health information systems for the Spanish National Health System (SNHS). A few certified data sources allow nation-wide public health monitoring and research.

The Primary Care Dataset [BDCAP] collects clinical data from a random sample of primary care episodes including information on active health problems, interventions, and intermediate health outcomes. BDCAP aims at studying the effectiveness, quality, and cost of primary care. BDCAP is curated and maintained by the MSSSI and it manages 2.79 million electronic health records (EHRs).¹

Primary Care Drugs Prescription Dataset [BIFAP] aims at carrying out pharmaco-epidemiological research in real life conditions. The information is voluntarily collected by physicians working in primary care settings. BIFAP includes clinical and prescription data from 4.8 million patients. BIFAP is fostered by the Spanish Agency of Medical Products and Devices, a public agency of the MSSSI.²

The Atlas of Variations in Medical Practice (Atlas VPM) collates all hospital admission data since 2002 in the SNHS, along with demographic, socioeconomic, and supply data. AtlasVPM aims at eliciting unwarranted differences in healthcare performance (i.e., effectiveness, quality and safety, and efficiency) across all SNHS healthcare providers. Atlas VPM, who currently manages 70 million hospital episodes, is curated and maintained by the Institute for Health Sciences in Aragon (IACS), a public institution of the SNS at regional level. All the research

¹ Details on BDCAP outlets can be found at: <https://www.msssi.gob.es/en/estadEstudios/estadisticas/estadisticas/estMinisterio/SIAP/home.htm> (accessed, July 2017)

² A detail of the projects developed in BIFAP might be found at <http://www.bifap.org/proyectos.php> (accessed, July 2017)

outputs are available in open access at www.atlasvpm.org.³

Taking advantage of the widespread coverage of EHR in Spain⁴, several ACs are developing major initiatives on their reuse for public health research and monitoring, e.g., BIGAN in Aragon or PADRI in Catalonia. Nevertheless, achievements are uneven across the SNHS, partly because of the concerns raised by pioneer initiatives on the inadequate use of EHR (which translated into stricter legal and ethical requirements), partly because of the lack of technical and methodological capacity.

3.6 Germany

The development of health care research in Germany has been late as compared to other countries. Nonetheless, it is supported by considerable funding by the BMBF (Ministry of education and research), institutions like the German Research Foundation, and the industrial sector. Many research centres for public health research were created with funds from BMBF leading to an increased number of high quality publications, e.g., [38], a development that was accelerated by the use of data regularly documented by statutory health insurances for research.

Health data providers include health insurance funds, institutes like the German Institute of Medical Documentation and Information (DIMDI), and different research facilities. Gaining access to data for research purposes may vary and be performed either per request, use contract, or project-specific agreements. The *GKV-Versorgungsstrukturgesetz* built the legal basis for the reuse of routine data collected by the statutory insurances (GKV) at DIMDI. DIMDI's Information System for Health Care Data started its pilot operation in 2014, and its data have been used in several

research projects. Data collected from health insurance companies directly contributed to the German Pharmacoepidemiological Research Database (GePaRD), whereby local health insurance data could be linked to death certificate data [39].

The German Network of Health Care Research has compiled best practices for health care research [40] in a series of memoranda that had a considerable influence on the research community. However, German public health research has still a lot to gain from initiatives to establish a modern Information Technology (IT) infrastructure for combining resources and enabling research with health routine data.

In summary, although we focused on a limited number of countries, we showed that health administrative data is abundant in Europe but access to this data for researchers is constrained by complexity and heterogeneity.

4 FAIR Principles: Impact for Public Health Informatics

Recently, a set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—has established a set of principles that we refer to as the FAIR Data Principles: data must be Findable, Accessible, Interoperable, and Reusable [41]. The FAIR principles put a specific emphasis on enhancing the ability of computers to automatically find and use the data, not only on supporting its reuse by individuals. Behind FAIR principles is the notion that algorithms are used to search for relevant data sources, to analyse the data sets, and to mine the data for scientific discovery. Therefore, the principles apply not only to the ‘data’, but also to the algorithms, tools, and workflows. The public health community will benefit from the application of the FAIR principles for:

- Data discovery (know which data is available and how to access the data),
- Accessing to aggregated data at the European level,
- Accessing to individual data, based on a secure environment respecting national and European policies on data protection,
- Services for hypothesis-driven studies through shared models and algorithms,

- Services to enable knowledge discovery with big data approaches.

Moreover, to enable data reuse, the public health informatics community must pave the road and provide (i) codes of practice and (ii) semantic frameworks that sources and users should agree upon in a near future.

5 Legal Framework

The EU legal framework consists of Regulations, Directives, Decisions, Recommendations and Opinions. The aim of the EU legal framework for public health is defined in a single statement by the EU Treaty of Lisbon [42]: “*A high level of human health protection shall be ensured in the definition and implementation of all Union policies and activities.*” This rule means that public health aspects should be included in all EU policies. This objective is to be achieved through community support to the member States. However, the regulation of healthcare systems continues to be the responsibility of the member states, and there exists a patchwork of different laws and policies [43], and a multitude of Health Acts, Social Security Funding Acts, Hospital Laws, Cancer Acts, Human Tissue Acts, Data Protection Acts, Health and Medicines Acts, Codes of Professional Conduct, and Codes of Medical Ethics. This patchwork of laws has created challenges for data sharing across jurisdictional borders. The third EU Health Programme (2014-2020) aims at supporting and adding value to the policies of Member States to improve the health of EU citizens, reduce health inequalities, and protect citizens from cross-border health threats (Reg. 282/2014). These efforts are supported by the regulation on European statistics on health (Reg. 1338/2008, 2015/359), and the establishment of a European network for the epidemiological surveillance and control of communicable diseases (Dec. 1082/2013, 2014/504, 2002/253).

The European Commission is unifying personal data protection with the General Data Protection Regulation (GDPR) [44], which will replace the Data Protection Directive 95/46/EC in 2018 thus imposing new obligations on organizations that process

³ Bernal-Delgado E, García-Armesto S, Peiró S on behalf of the Atlas VPM group. Atlas of Variations in Medical Practice: the Spanish National Health Service under scrutiny. *Health Policy* 2014;114(1):15-30.

⁴ OECD/EU (2016), *Health at a Glance: Europe 2016: State of Health in the EU Cycle*, OECD Publishing, Paris. <http://dx.doi.org/10.1787/9789264265592-en>

personal data of EU residents, with a major impact on research involving health data [57]. The GDPR is an attempt to reconcile the often competing requirements of privacy protection and innovations in data management: (i) organizations that process personal health data for research purposes have to implement appropriate safeguards; (ii) under certain conditions, they may override a subject's right to object to the processing of his/her data or to seek the erasure of his/her personal data (Art. 89). It is an exemption to the principle of purpose limitation for research through Art. 5(1)(b), 89(1), and 9(2)(j).

Public health research is treated as a subset of scientific research (Rec. 159). Therefore, the same exemptions and requirements apply here. However, the GDPR also contains several provisions applicable exclusively to public health research and covering (i) higher protections for the processing of sensitive data for health-related purposes; (ii) transfer of personal data to third countries if transfer is necessary for important reasons of public interest, e.g., contagious diseases; (iii) stronger requirements to consult supervisory authorities about processing activities.

Public health research must pay special attention to legal issues [45], e.g., a registry data owner's country legislation may forbid cross-border exchange of specific data. Moreover, a risk assessment has to be considered concerning the risk of re-identification from individual data. This risk depends on the nature and availability of contextual information, and also on IT capabilities [46]. Unlike the United States Health Insurance Portability and Accountability Act (HIPAA), which exempts data from regulation if 18 specific personal identifiers are removed, GDPR considers data as anonymous only when it cannot be identified by any means "reasonably likely to be used (...) either by the controller or by any other person" (Rec. 23). Thus, even if a user is neither able nor willing to re-identify individual data, a data set may still fall under the GDPR if it could be re-identified "with reasonable effort". Public health research must build and maintain trust and this can only be done when employing the best state-of-the-art IT techniques for the protection of sensitive data against misuse.

6 Discussion and Conclusion

Funding and governmental agencies require adequate data management plans for data generated by publicly funded research. Beyond proper collection, annotation, and archival, data stewardship includes the notion of FAIR data, with the goal that the data should be available to be re-used for downstream investigations [41]. Conditions that simplify discovery, evaluation, and reuse include the ability to implement shared metadata, algorithms, and services. Until now, most European projects in public health have developed their own, mainly short-term, project-specific solutions resulting in the need for addressing more far-reaching interoperability and data sharing issues.

The notion of reference datasets already exists for certain domains, like genomics. Some efforts have been made to deliver such data sets in public health, e.g., the French "Echantillon Généraliste des Bénéficiaires" is an anonymous sample extracted from SNIIRAM. However, conversely to standardized high-throughput data in genomics, public health data is mainly made of heterogeneous and low-throughput data. The resulting ecosystem for health data might become more diverse, thereby exacerbating the discovery, interoperability, and re-usability issues, not only by humans, but also through the automatic processing by computers. In this article, we put an emphasis on health administrative databases, which bring even more difficulties.

Following the example of the Nordic countries, several European countries aim at facilitating the re-use of their health administrative databases for research. This will increase scientific knowledge in a number of areas where the power of large databases is needed. However, only sophisticated IT methods can ensure appropriate data protection in that case.

Most of the advances in public health are being driven when data of high quality, combining population genomics, routine care data, biobank annotations, costs, and other sources, are combined to more traditional public health data. The convergence of several national initiatives in Europe, the adoption of the FAIR principles, the in-

creasing number of studies reusing routine data, and the experience gained by the medical informatics community in addressing interoperability issues, all are incentives to develop a common public health research infrastructure in Europe. An infrastructure able to provide secure and efficient access to health administrative data would provide without doubt a boost for the research capacities of existing bio-medically-focused research infrastructures.

References

1. Pisani E, Whitworth J, Zaba B, Abou-Zahr C. Time for fair trade in research data. *Lancet* 2010 Feb 27;375(9716):703-5.
2. Chêne G, Phillips A, Costagliola D, Sterne JA, Furrer H, Del Amo J, et al. Cohort Profile: Collaboration of Observational HIV Epidemiological Research Europe (COHERE) in EuroCoord. *Int J Epidemiol* 2016 Nov 17. pii: dyw211.
3. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. *Bull World Health Organ* 2010 Jun;88(6):462-6.
4. Ainsworth J, Buchan I. Combining Health Data Uses to Ignite Health System Learning. *Methods Inf Med* 2015;54(6):479-87.
5. Ethier JF, Curcin V, Barton A, McGilchrist MM, Bastiaens H, Andreasson A, et al. Clinical data integration model. Core interoperability ontology for research using primary care data. *Methods Inf Med* 2015;54(1):16-23.
6. Roberto G, Leal I, Sattar N, Loomis AK, Avillach P, Egger P, et al. Identifying cases of Type 2 Diabetes in heterogeneous data sources: strategy from the EMIF project. *PLoS One* 2016 Aug 31;11(8):e0160648.
7. Delnord M, Hindori-Mohangoo AD, Smith LK, Szamotulska K, Richards JL, Deb-Rinker P, et al. Variations in very preterm birth rates in 30 high-income countries: are valid international comparisons possible using routine data? *BJOG* 2017 Apr;124(5):785-94.
8. Bernal-Delgado E, Christiansen T, Bloor K, Mateus C, Yazbeck AM, Munck J, et al. ECHO: health care performance assessment in several European health systems. *Eur J Public Health* 2015 Feb;25 Suppl 1:3-7.
9. Heijink R, Engelfriet P, Rehnberg C, Kittelsen SA, Häkkinen U, et al. A Window on Geographic Variation in Health Care: Insights from EuroHOPE. *Health Econ* 2015 Dec;24 Suppl 2:164-77.
10. Mari-Dell'Olmo M, Gotsens M, Palència L, Burström B, Corman D, Costa G, et al. Socioeconomic inequalities in cause-specific mortality in 15 European cities. *J Epidemiol Community Health* 2015 May;69(5):432-41.
11. Avillach P, Coloma PM, Gini R, Schuemie M, Mougou F, Dufour J-C, et al. Harmonization process for the identification of medical events in eight European healthcare databases: the experi-

- ence from the EU-ADR project. *J Am Med Inform Assoc* 2013 Jan 1;20(1):184–92.
12. Charlton RA, Neville AJ, Jordan S, Pierini A, Damase-Michel C, Klungsoyr K, et al. Healthcare databases in Europe for studying medicine use and safety during pregnancy. *Pharmacoepidemiol Drug Saf* 2014 Jun;23(6):586–94.
 13. Reynolds RF, Kurz X, de Groot MCH, Schlienger RG, Grimaldi-Bensouda L, Tcherny-Lessenot S, et al. The IMI PROTECT project: purpose, organizational structure, and procedures. *Pharmacoepidemiol Drug Saf* 2016 Mar;25 Suppl 1:5–10.
 14. Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, et al. Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies. *EGEMS (WashDC)* 2016 Feb 8;4(1):1189.
 15. Andersen JS, Olivarius Nde F, Krasnik A. The Danish National Health Service Register. *Scand J Public Health* 2011 Jul;39(7 Suppl):34–7.
 16. Ludvigsson JF, Almqvist C, Bonamy AK, Ljung R, Michaëlsson K, Neovius M, et al. Registers of the Swedish total population and their use in medical research. *Eur J Epidemiol* 2016 Feb;31(2):125–36.
 17. Berger SM, Gislason G, Moore LL, Andersson C, Torp-Pedersen C, Denis GV, et al. Associations between metabolic disorders and risk of cancer in Danish men and women—a nationwide cohort study. *BMC Cancer* 2016 Feb 22;16:133.
 18. Bjerrum A, Andersen O, Fischer A, Lindebjerg J, Lynge E. Colorectal cancer mortality 10 years after a single round of guaiac faecal occult blood test (gFOBT) screening: experiences from a Danish screening cohort. *BMJ Open Gastroenterol* 2016 Dec 7;3(1):e000120.
 19. Goldacre B. The NHS plan to share our medical data can save lives – but must be done right. *The Guardian*, 21 February 2014.
 20. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data ran into trouble. *J Med Ethics* 2015;41:404–9.
 21. Herrett E, Gallagher AM, Bhaskaran K, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;44:827–36.
 22. O'Meara H, Carr DF, Evelyn J, Hobbs M, McCann G, van Staa T, et al. Electronic health records for biological sample collection: feasibility study of statin-induced myopathy using the Clinical Practice Research Datalink. *Br J Clin Pharmacol* 2014;77:831–8.
 23. Dregan A, van Staa T, McDermott L, McCann G, Ashworth M, Charlton J, et al. Point-of-care cluster randomized trial in stroke secondary prevention using electronic health records. *Stroke* 2014 Jul;45(7):2066–71.
 24. van Staa T, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B, et al. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health Technol Assess* 2014;18:1–146.
 25. Jones KH, Ford DV, Jones C, Dsilva R, Thompson S, Brooks CJ, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: a privacy-protecting remote access system for health-related research and evaluation. *J Biomed Inform* 2014;50:196–204.
 26. Scottish Health Informatics Programme. http://www.scot-ship.ac.uk/sites/default/files/Reports/SHIP_BLUEPRINT_DOCUMENT_final_100712.pdf (accessed 18 Feb 2016).
 27. van Staa TP, Goldacre B, Buchan I, Smeeth L. Big health data: the need to earn public trust. *BMJ* 2016 Jul 14;354:i3636.
 28. Duong M, Gulmez SE, Salvo F, Abouelfath A, Lassalle R, Droz C, et al. Usage patterns of paracetamol in France. *Br J Clin Pharmacol* 2016 Aug;82(2):498–503.
 29. Fautrel B, Cukierman G, Joubert JM, Laurendeau C, Gourmelen J, Fagnani F. Healthcare service utilisation costs attributable to rheumatoid arthritis in France: Analysis of a representative national claims database. *Joint Bone Spine* 2016 Jan;83(1):53–6.
 30. Filipovic-Pierucci A, Samson S, Fagot JP, Fagot-Campagna A. Estimating the prevalence of depression associated with healthcare use in France using administrative databases. *BMC Psychiatry* 2017 Jan 3;17(1):1.
 31. Massoulié G, Wintzer-Wehekind J, Chenaf C, Mulliez A, Pereira B, Authier N, et al. Prognosis and management of myocardial infarction: Comparisons between the French FAST-MI 2010 registry and the French public health database. *Arch Cardiovasc Dis* 2016 May;109(5):303–10.
 32. Gini R, Schuemie MJ, Mazzaglia G, Lapi F, Francesconi P, Pasqua A, et al. Automatic identification of type 2 diabetes, hypertension, ischaemic heart disease, heart failure and their levels of severity from Italian General Practitioners' electronic medical records: a validation study. *BMJ Open* 2016 Dec 1;6(12):e012413.
 33. Mazzali C, Duca P. Use of administrative data in healthcare research. *Intern Emerg Med* 2015 Jun;10(4):517–24.
 34. Benedetti MD, Pugliatti M, D'Alessandro R, Beghi E, Chiò A, Logroscino G, et al. A Multicentric Prospective Incidence Study of Guillain-Barré Syndrome in Italy. The ITANG Study. *Neuroepidemiology* 2015;45(2):90–9.
 35. Giner-Soriano M, Roso-Llorach A, Vedia Urgell C, Castells X, Capellà D, Ferreira-González I, et al. Effectiveness and safety of drugs used for stroke prevention in a cohort of non-valvular atrial fibrillation patients from a primary care electronic database. *Pharmacoepidemiol Drug Saf* 2017 Jan;26(1):97–107.
 36. Huerta C, Abbing-Karahagopian V, Requena G, Oliva B, Alvarez. Exposure to benzodiazepines (anxiolytics, hypnotics and related drugs) in seven European electronic healthcare databases: a cross-national descriptive study from the PROTECT-EU Project. *Pharmacoepidemiol Drug Saf* 2016;25:56–65.
 37. Angulo-Pueyo E, García-Armesto S, Martínez-Lizaga N, Riado-López M, Peiró S, Bernal-Delgado E. Factors associated with hospitalisations in chronic conditions deemed avoidable: ecological study in the Spanish health care system. *BMJ Open* 2017 Feb 24;7(2):e011844.
 38. Hajek A, Bock JO, König HH. Which factors affect health care use among older Germans? Results of the German ageing survey. *BMC Health Serv Res* 2017 Jan 13;17(1):30.
 39. Ohlmeier C, Langner I, Garbe E, Riedel O. Validating mortality in the German Pharmacoepidemiological Research Database (GePaRD) against a mortality registry. *Pharmacoepidemiol Drug Saf* 2016 Jul;25(7):778–84.
 40. Müller D, Augustin M, Banik N, Baumann W, Bestehorn K, Kieschke J, et al. [Memorandum registry for health services research]. [Article in German]. *Gesundheitswesen*. 2010 Nov;72(11):824–39.
 41. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016 Mar 15;3:160018.
 42. Consolidated version of the Treaty on the Functioning of the European Union – Part 3: Union Policies and Internal Actions - Title XIV: Public Health - Article 168, Official Journal L115, 09/05/2008. p. 122–4.
 43. Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf
 44. Maldoff G. How GDPR changes the rules for research. *The Privacy Advisor* Apr 19, 2016.
 45. Lowrance WW. Privacy, Confidentiality, and Health Research. Cambridge University Press, Cambridge UK; 2012.
 46. Kuchinke W, Ohmann C, Verheij RA, van Veen E-B, Arvanitis TN, Taweel A, et al. A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model. *Int J Med Inform* 2014 Dec;83(12):941–57.

Correspondence to:

Anita Burgun
 Inserm, UMR 1138
 Centre de Recherche des Cordeliers, AP-HP
 Paris Descartes University
 Paris, France
 E-mail: anita.burgun@aphp.fr