

# Scaling up health knowledge at European level requires sharing integrated data: an approach for collection of database specification

Enrica Menditto<sup>1</sup>  
 Angela Bolufer De Gea<sup>2</sup>  
 Caitriona Cahir<sup>3,4</sup>  
 Alessandra Marengoni<sup>5</sup>  
 Salvatore Riegler<sup>1</sup>  
 Giuseppe Fico<sup>6</sup>  
 Elisio Costa<sup>7</sup>  
 Alessandro Monaco<sup>8</sup>  
 Sergio Pecorelli<sup>5</sup>  
 Luca Pani<sup>8</sup>  
 Alexandra Prados-Torres<sup>9</sup>

<sup>1</sup>School of Pharmacy, CIRFF/Center of Pharmacoeconomics, University of Naples Federico II, Naples, Italy; <sup>2</sup>Directorate-General for Health and Food Safety, European Commission, Brussels, Belgium; <sup>3</sup>Division of Population Health Sciences, Royal College of Surgeons in Ireland, <sup>4</sup>Department of Pharmacology and Therapeutics, St James's Hospital, Dublin, Ireland; <sup>5</sup>Department of Clinical and Experimental Science, University of Brescia, Brescia; <sup>6</sup>Life Supporting Technologies, Photonics Technology and Bioengineering Department, School of Telecommunications Engineering, Polytechnic University of Madrid, Madrid, Spain; <sup>7</sup>Faculty of Pharmacy, University of Porto, Porto, Portugal; <sup>8</sup>Italian Medicines Agency – AIFA, Rome, Italy; <sup>9</sup>EpiChron Research Group on Chronic Diseases, Aragón Health Sciences Institute (IACS), IIS Aragón REDISSEC ISCIII, Miguel Servet University Hospital, University of Zaragoza, Zaragoza, Spain

Correspondence: Enrica Menditto  
 School of Pharmacy, CIRFF/Center of Pharmacoeconomics, University of Naples Federico II, Via Domenico Montesano, 49, 80131 Naples, Italy  
 Tel +39 81 67 8660  
 Fax +39 81 67 8303  
 Email enrica.menditto@unina.it

**Abstract:** Computerized health care databases have been widely described as an excellent opportunity for research. The availability of “big data” has brought about a wave of innovation in projects when conducting health services research. Most of the available secondary data sources are restricted to the geographical scope of a given country and present heterogeneous structure and content. Under the umbrella of the European Innovation Partnership on Active and Healthy Ageing, collaborative work conducted by the partners of the group on “adherence to prescription and medical plans” identified the use of observational and large-population databases to monitor medication-taking behavior in the elderly. This article describes the methodology used to gather the information from available databases among the Adherence Action Group partners with the aim of improving data sharing on a European level. A total of six databases belonging to three different European countries (Spain, Republic of Ireland, and Italy) were included in the analysis. Preliminary results suggest that there are some similarities. However, these results should be applied in different contexts and European countries, supporting the idea that large European studies should be designed in order to get the most of already available databases.

**Keywords:** health care databases, adherence, electronic health records, outcome research

## Background

The rise of critical questions on health outcomes, effectiveness, and impact of medical plans and therapies on older adults over the last decades has led to an exploration of new methodologies and research approaches.<sup>1</sup> Computerized health care databases have been widely described as an excellent opportunity for secondary data used in research.<sup>2,3</sup> These databases include electronic health records and administrative data already collected from large populations for other purposes, such as hospital discharge, prescribed drugs, and procedures. They can be subsequently merged at an individual level using unique, anonymized identifiers, making data available for both academic and policy research.<sup>4,5</sup> Big data, defined as “massive, complex, distributed, and often dynamic sets of data”,<sup>6</sup> are generated from increasingly diverse sources and offer opportunities to better understand known trends and to discover new ones as well as relationships that were indiscernible until now. This is due to the “4Vs” (velocity, variety, volume, and especially veracity), which in turn depend on data quality and assurance. The availability of big data has brought about a wave of innovation in projects when conducting health services research,<sup>6</sup> thus adding semantic capabilities by enriching and tagging anonymized documentation systems automatically fed by public health care registries. In these studies, patients can be profiled on the basis of



the potential risk of diverse health outcomes (eg, hospital admission, adverse drug events). All this requires advanced computational frameworks for high data volume and intensive data processing.

The advantages of large-population databases linked at the patient level are their large dimension (often the whole population of a given country or region), data on real-life use, and outcomes and detection of long-term effects that are not observable through randomized controlled trials. On the other hand, a significant characteristic of anonymized repositories of health databases is reusability. Indeed, while collection of patient information during trials is designed to gather information on a given subject in a specific time frame, large-population databases aim to track events over patients' lifetimes, such as drug prescription or hospitalization, thus enabling multiple information combinations that can be used in many different applications. Linked health electronic databases represent powerful and relatively low-cost resources for investigating important public health concerns in real-life scenarios covering large populations.<sup>7,8</sup>

Several institutions in Europe are conducting research in this field exploring appropriateness of drug use in the elderly,<sup>9,10</sup> adherence to therapy,<sup>11,12</sup> polypharmacy patterns, and use of health resources,<sup>13,14</sup> but typically as isolated bodies. Most of the available secondary data sources are restricted to the geographical scope of a given country and present heterogeneous structure and content (for example, type of collected data, drug, and clinical event terminologies) even if trends in the recent years have shown that the number of studies conducted using multiple databases is on the gradual increase.<sup>15</sup>

Using multiple data sources is not an easy task, as it implies a set of multiple actions to be taken, such as data and meta-data analysis, identification of common data sets, solving privacy and data property issues, and data integration.

Under the umbrella of the European Innovation Partnership on Active and Healthy Ageing (EIP on AHA), collaborative work conducted by the partners of the specific group (the Adherence Action Group) on "adherence to prescription and medical plans" identified the use of observational databases related to large populations to monitor medication-taking behavior in the elderly. A collection of characteristics of databases was devised to collaborate in building consensus on the use of common elements and measures to facilitate data sharing. This article goes into detail in a descriptive manner on the methodology used to gather the information from the available databases among

the Adherence Action Group partners with the purpose of paving the way toward improved data sharing in order to conduct research on a European level by providing a framework to undertake data sharing.

## Methods

### European Innovation Partnership on Active and Healthy Ageing

Under the Innovation Union<sup>16</sup> flagship initiative of the Europe 2020 Strategy,<sup>17</sup> the EIP on AHA was set up as the first European innovation public-private partnership bringing together all the relevant actors of the European Union (EU), at national and regional levels, in order to tackle barriers to innovation for the societal challenge that active and healthy aging represents. The EIP on AHA aims to identify and remove persistent barriers to innovation for active and healthy aging through interdisciplinary and cross-sectorial approaches focused on improving the health and quality of life of Europeans, especially older people. The idea underpinning the EIP on AHA is to support the long-term sustainability and efficiency of health and social care systems and to enhance the competitiveness of EU industry through business and expansion in new markets.

Under the pillar of "prevention, screening, and early diagnosis", the EIP on AHA identified the priority area on "adherence to medication and medical plans" (the Adherence Action Group) in order to deliver tangible adherence approaches for patients in various disease areas, at the regional level and in different member states. The Adherence Action Group brings together partners representing 68 multistakeholder commitments from the national, regional, and local authorities; research centers; academia; industry; enterprises; and existing consortiums across the EU.<sup>18</sup> The work of this group focuses on improving patient adherence to medical plans, empowering patients and care givers to take care of their health and to be independent, delivering improvements in the health care systems, improving existing data evidence on aging and adherence, and enhancing communication between different actors in the healing and caring process. As part of the collaborative work conducted in the group, partners identified the use of observational and large-population databases as a tool to carry out evidence on medication-taking behavior in the elderly. To scale up the results individually achieved in each country, it is necessary to create a multiheaded network of databases sharing a common structure in order to identify a minimum common data set that can be freely used by all partners joining the network.

The process described here comprises two phases: 1) collection and characterization of electronic databases and 2) measurement of adherence in the older population.

## Collection and characterization of electronic databases

First, we provided a description of all individual health-related databases available to the partners taking part in the Adherence Action Group in terms of scope, structure, content, data fields, and records. We also identified common data elements (CDEs). The different data countries were surveyed.

For this purpose, a computer engineer (SR) had set up a document containing a total of 23 categorical questions regarding the scope of database, structure, content, and data sources to describe database metadata (Figure S1). This document was sent in July 2014 to all interested partners who answered by filling in as many questionnaires as databases owned. In case of information over or under detailed, the documents were adjusted to fit in the overall document structure. Two domain experts (AM, AM) validated the provided data. Finally, two domain experts (EM, AP-T) analyzed the data collected via email. Partners involved in the data collection process reviewed the results and contributed to the final manuscript.

## Measurement of adherence in the older population

Once the different databases were identified and analyzed in terms of their characteristics, we collected information on research initiatives, in terms of study methods and preliminary results, in which the Adherence Action Group partners were involved. More in detail, we focused on results regarding the measurement of adherence in the older population through observational studies and based on the already described databases. To this end, each partner provided, through a form allowing both multiple choice and free text, detailed essential information on planned or ongoing research initiatives related to adherence in the older population (Figure S2). In those cases in which the information provided by partners was over or under detailed, small adjustments were made in order to adapt provided data to the overall document structure. In the second stage, information provided by each research group was analyzed in the light of the specificities of each study and organized according to the studied aspects related to adherence (ie, objective of the study, data source, population and time frame, type of medication, and operational definition of

adherence indicator, other variables, and outcomes studied). Last, research groups were asked to provide, if appropriate, preliminary results of respective ongoing studies, key findings of which were extracted and summarized. The first form was circulated in September 2014 to partners participating in the Adherence Action Group. The form was sent, and the results were collected via email.

## Results

A total of six databases belonging to three different European countries were included in the analysis: the EpiChron Integrated Health Database (EpiChron-IHD) from Spain, The Irish Longitudinal study on Ageing (TILDA) from the Republic of Ireland, the Optimizing Prescription in Elderly in Nursing Home from Italy, the L'Osservatorio Nazionale sull'Impiego dei Medicinali from Italy, the Campania Region Chronic Disease Analysis from Italy, and the Caserta Health Unit Administrative Medication Data Warehouse from Italy (Table 1).

The data sources varied from country to country depending on the level of sophistication of record keeping; data collection, analysis, and reporting; and the operational considerations of the health care system. Four of the databases described had electronic health record databases as data sources established for routine administrative purposes (ie, EpiChron-IHD, L'Osservatorio Nazionale sull'Impiego dei Medicinali, Campania Region Chronic Disease Analysis, and Caserta Health Unit Administrative Medication Data Warehouse), one was related to registry data (Optimizing Prescription in Elderly in Nursing Home) and one integrated administrative data sources with patient-reported data (TILDA-Health Services Executive Primary Care Reimbursement Service [HSE-PCRS]). In particular, for the TILDA-HSE-PCRS database, data were linked for 3,122 older community dwelling participants to the HSE-PCRS pharmacy claims database that contained details on all drugs dispensed. Of note, EpiChron-IHD also integrated clinical information from general practitioners' medical records (Figure 1).

In all databases, drugs were coded using the Anatomical Therapeutic Chemical classification. Diagnoses were coded using International Classification of Diseases, Ninth Revision, Clinical Modification classification in all databases except one where it was coded by using International Classification of Primary Care (Table 2).

The databases were specifically used, in the context of the EIP on AHA, to carry out studies on multimorbidity, polypharmacy, and medication-taking behaviors.

**Table I** Characteristics of the databases

Database	EpiChron-IHD	TILDA	OPEN	OsMed Health-DB Database	CaRe_CroDA	CE_AdMeDa
Responsible organization	Aragón Health Sciences Institute (IACS)/ EpiChron Research Group on Chronic Diseases	Trinity College Dublin EngAGE, the Centre for Research in Ageing	AIFA, Italian Medicines Agency; University of Brescia; and Mario Negri Institute, Milan	CliCon under commission by AIFA for OsMed Survey	CIRFF/Center of Pharmacoconomics, University of Naples Federico II	CIRFF/Center of Pharmacoconomics, University of Naples Federico II, under commission by Local Health Unit of Caserta
Geographic area	Aragon, Spain	The Republic of Ireland	Brescia, Northern Italy	Italy	Campania, Italy	Caserta, Southern Italy
Population covered	1,300,000	1,500,000	236,000	29,000,000	950,000	908,000
Population covered by database	1,270,000	8,000	700	29,000,000 (6,100,000 65+ years old)	950,000	725,000
Age span covered by database	Whole population	50+ years old	65+ years old	Whole population	65+ years old	Whole population
System category	Major application	General support system	General support system	General support system	Major application	General support system
Time span covered	2010–2011	2009–2014	2013–2014	2009–2014	2009–2011	2009–2014
Scope	Multimorbidity, patterns of chronic diseases and their relation with prescription profiles, quality of care, use of health services, and pharmacoepidemiology, including adverse drug events	To assess MTB and relationship with health outcomes for drugs prescribed for different conditions	Prescription appropriateness in the elderly residents in nursing homes	Disease prevalence/ incidence studies, drug utilization studies, health outcome studies, studies on the use of health resources, appropriateness and adherence analyses, and qualitative performance indicators	To assess MTB and relationship with health outcomes in common chronic conditions	Pharmacoepidemiologic and pharmaco-economic analyses

**Abbreviations:** EpiChron-IHD, EpiChron Integrated Health Database; TILDA, The Irish Longitudinal study on Ageing; OPEN, Optimizing Prescription in Elderly in Nursing Home; OsMed, L'Osservatorio Nazionale sull'Impiego dei Medicinali; CaRe\_CroDA, Campania Region Chronic Disease Analysis; CE\_AdMeDa, Caserta Health Unit Administrative Medication Data Warehouse; MTB, medication-taking behavior; DB, database; IACS, Aragón Health Sciences Institute; CIRFF, Center of Pharmacoconomics; AIFA, Italian Drug Agency.

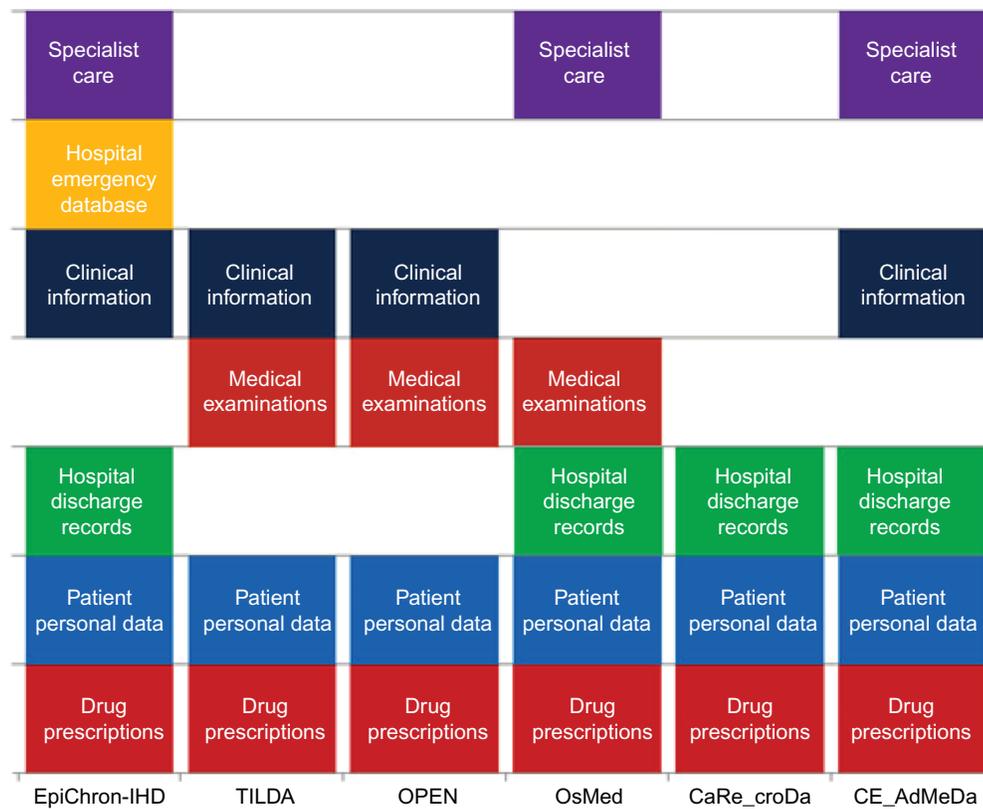
Regarding the measurement of adherence in the older population, one of the main findings was the heterogeneity of the study designs applied. Although all research initiatives were observational, some of them specifically focused on potential determinants of nonadherence, whereas others on consequences and health outcomes (eg, nonadherence, risk of hospitalization, quality of life). Some studies focused on specific medications (eg, antiosteoporotic treatment, antidepressants) and others on specific chronic diseases (eg, ischemic heart disease, neurological diseases, or mental diseases). All these studies were related to chronic diseases. Medication possession ratio and/or proportion of days covered were frequently used as indicators to measure adherence and persistence to medication. No studies reported adherence to aspects related to medical care besides medication, such as physician office visit and patient monitoring.

Regarding sociodemographic characteristics, sex and age information was included in all studies. Comorbidity and polypharmacy were also variables considered in most studies. Although most of them were still ongoing, preliminary results suggested some common findings. For example, nonadherence was more prevalent among younger people and highly influenced by the concomitant presence of mental health problems. Furthermore, comorbidity and complexity of drug regimens were positively associated with nonadherence.

## Discussion

### Principal findings

Our analysis showed that health care databases available to the partners involved in the Adherence Action Group rely



**Figure 1** Specific data sources contributing to each database.

**Abbreviations:** EpiChron-IHD, EpiChron Integrated Health Database; TILDA, The Irish Longitudinal study on Ageing; OPEN, Optimizing Prescription in Elderly in Nursing Home; OsMed, L'Osservatorio Nazionale sull'Impiego dei Medicinali; CaRe\_CroDA, Campania Region Chronic Disease Analysis; CE\_AdMeDa, Caserta Health Unit Administrative Medication Data Warehouse.

on different technologies and have specific data structure designs. However, these issues can be overcome by implementing data translation layers toward a common structure, in order to exploit the potential that current advances in technology could provide.

Despite complexity, combining databases from different countries, although a challenging task, is possible. Combining databases from multiple countries exploiting common structural elements will help increasing the cohorts both on numerical and geographical coverage aspects. To allow this, infrastructures should allow for data identification, query of data, merging data from different sources, and transference of data following security and privacy guidelines. They should deal with different databases located across different countries and allow access to third countries. They should also grant integrity and security of the data transferred, allow to run cross-queries across different databases, and inform about the precedence of the data provided as results. In order to enable secondary use of health care data and bridge the gap between clinical and research domains, several initiatives have been carried out.

Patient care and research need and use different data models. Accordingly, CDE models have been developed, such as C154 – Data Dictionary Component, the Federal Health Information Model, and the domain analysis model.<sup>20–23</sup> Another useful approach aims to define both CDEs and accompanying data models. An example is Informatics for Integrating Biology and the Bedside (i2b2) designed for cohort identification. Many of the queries asked in a registry are essentially forms of cohort identification (eg, how many patients are on medication X, how many are adhering to evidence-based guideline Y). In addition, building registries on top of i2b2 removes the need to either: 1) load the data into multiple database systems or 2) have users manually reenter the relevant electronic medical record data. By building research registries using i2b2, users can add data that are not collected in the electronic medical record.<sup>24</sup>

Our study indicated that the databases considered are already being used for analysis in the field of multimorbidity, adherence to medication, and polypharmacy. Although most of these studies are still ongoing, preliminary results suggest

**Table 2** Specific fields in the data sources contributing to each database

Data sources	EpiChron-IHD	TILDA	OPEN	OsMed	CaRe_CroDA	CE_AdMeDa
	Patient ID Sex Date of birth	Field group: Demographic data Socioeconomic data General health-related data Social support and intergenerational transfers Personal health behaviors	Patient ID Sex Date of birth	Patient ID Sex Date of birth	Patient ID Sex Date of birth	Patient ID Sex Date of birth
Drug prescriptions	Patient ID Drug code Prescription_date Delivery_date Quantity ATC code Price DDD	Field group: Self-reported current medications Drug dispensing history from linked pharmacy claims data	Field group: Drug prescriptions	Patient ID Drug code Prescription_date Delivery_date Quantity ATC code Price DDD	Patient ID Drug code Prescription_date Delivery_date Quantity ATC code Price DDD	Patient ID Drug code Prescription_date Delivery_date Quantity ATC code Price DDD
Hospital discharge records	Patient ID Type of admission Date of admission Reasons for discharge Diagnoses (ICD-9) Date of discharge			Patient ID Type of admission Date of admission Reasons for discharge Diagnoses (ICD-9 code) Procedures (ICD-9 code) Date of discharge DRG	Patient ID Type of admission Date of admission Reasons for discharge Diagnoses (ICD-9 code) Date of discharge	Patient id Type of admission Date of admission Reasons for discharge Diagnoses (ICD-9 code) Date of discharge
Medical examinations		Field group: Health assessment	Field group: Medical examinations	Patient ID Date Type of test Result (available on 10% of the sample)		
Clinical information from primary care	Patient ID Diagnoses (ICPC) Diagnoses opening date Diagnoses closing date	Field group: Health assessment	Field group: Clinical examinations			Patient ID Diagnoses (ICD-9-CM) (available on 20% of the sample)
Hospital emergency database	Patient ID Diagnoses (ICD-9) Discharge date					
Specialist care	Patient ID Date Source of referral			Patient ID Date Type of test/visit Price		Patient ID Date Type of test/visit Price

**Abbreviations:** EpiChron-IHD, EpiChron Integrated Health Database; TILDA, The Irish Longitudinal study on Ageing; OPEN, Optimizing Prescription in Elderly in Nursing Home; OsMed, L'Osservatorio Nazionale sull'Impiego dei Medicinali; CaRe\_CroDA, Campania Region Chronic Disease Analysis; CE\_AdMeDa, Caserta Health Unit Administrative Medication Data Warehouse; ATC, Anatomical Therapeutic Chemical; ICD-9, International Classification of Diseases, Ninth Revision; ICPC, International Classification of Primary Care; ICD-9-CM, International Classification of Diseases, Ninth Revision, Clinical Modification; DRG, diseases related group; DDD, defined daily dose.

that there are common findings. For example, it seems that nonadherence is more prevalent among younger people. Furthermore, comorbidity and complexity of drug regimens are positively associated with nonadherence. These results

should be extended in different contexts and European countries, supporting the idea that large European studies should be designed in order to get optimal results out of already available databases. Furthermore, some public health issues

require more representative populations, longer follow-up periods, and a greater range of patient data.<sup>25</sup> These kinds of findings may be highly relevant when planning initiatives to increase adherence by more focused interventions.

## European studies combining data from multiple data sources

A recent survey carried out by the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance highlighted that multinational studies combining data from multiple sources have been encouraged over the recent years by funding calls of the European Commission such as Seventh Framework Programme, Horizon 2020 program, and the Innovative Medicines Initiative. Approximately 18 projects have been totally or partially publicly funded by the European Commission in the years 2008–2013, although the majority of them are focused on drug safety. The methodology used to combine data from multiple databases has been heterogeneous.<sup>26</sup> Among these projects, some examples are Exploring and Understanding Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge (EU-ADR)<sup>27</sup> and Pharmacoepidemiological Research on Outcomes of Therapeutics (PROTECT).<sup>28</sup> The EU-ADR project aims to design, develop, and validate a computerized integrative system for early detection of adverse drug reactions. This collaborative framework uses data extracted from eight European health records databases from four countries (Italy, the Netherlands, the UK, and Denmark) and comprises data of >30 million patients. The resulting platform was used to run specific web services and workflows and combine obtained evidence to constituting an effective tool to help research work in pharmacovigilance.<sup>29,30</sup> The PROTECT project aims to strengthen the monitoring of the benefit–risk of medicines in Europe, which consists of 34 public and private partners coordinated

by the European Medicines Agency. In this project, a methodological framework for pharmacoepidemiological studies for signal detection and evaluation in various types of datasets was developed and tested.<sup>31</sup> Another example is Survey of Health, Ageing and Retirement in Europe project, a multidisciplinary and cross-national panel database of micro data on health, social and family network, and socioeconomic status of those aged ≥50 years<sup>32,33</sup> from 20 European countries (+Israel; Table 3).

In addition to these initiatives, the European Commission and the WHO Regional Office for Europe agreed in 2010 to strengthen cooperation in order to work toward a single information system for health in Europe, building on existing cooperation and also expanding the use of shared data collection, collaborative analyses of health issues, and generation and dissemination of knowledge in support of health policy. The report “Promoting better integration of health information systems: best practices and challenges” published in 2015 addresses the current trends in the member states of the EU and European Free Trade Association on how to promote better integration of health information systems. To understand what better integration means from a pragmatic perspective, the Health Evidence Network conducted interviews with experts from 13 member states of the EU, the results of which were combined with the findings from a literature search. The results from the interviews stress the need 1) for ongoing work on some “basics”, such as data availability and quality, inventories of data and registries, standardization, legislation, physical infrastructure, and workforce capacities; 2) to continue with the work on more “concept-driven” indicator sets; 3) to define what better integration means and to demonstrate concrete benefits of integration; 4) to build leadership for capacity building in further integration of health information systems; and 5) for a further international exchange about ongoing activities in this area.<sup>34</sup>

**Table 3** List of EC-funded projects on database sharing

Project/Study	Funding calls	Outcome	No of databases integrated	Type of initiative
EU-ADR project <sup>27</sup>	FP7	Design, develop, and validate a computerized integrative system for early detection of adverse drug reactions	Eight European health record databases from four countries	Computerized integrated framework using EHR and biomedical data
PROTECT <sup>28</sup>	IMI	Early detection of adverse events	n/a	Methodological framework for pharmacoepidemiological studies for signal detection and evaluation
SHARE <sup>32</sup>	FP7	Survey of health, aging, and retirement in Europe	20 European countries (+ Israel)	Panel database of micro data on health, social and family network, and socioeconomic status

**Abbreviations:** EC, European Commission; EU-ADR, Exploring and Understanding Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge; FP7, Seventh Framework Programme; EHR, electronic health records; PROTECT, Pharmacoepidemiological Research on Outcomes of Therapeutics; IMI, Innovative Medicines Initiative; n/a, not applicable; SHARE, Survey of Health, Ageing and Retirement in Europe.

## Strengths and limitations

To our knowledge, this is the first initiative carried out at the European level to identify and compare health databases to study adherence to medical plans. To this end, the EIP on AHA initiative represents a unique opportunity to compare practices, identify common needs and gaps, and establish good practices and harmonized approaches with a view to maximize the effective exploitation of large data sets and provide the basis for studying population cohorts at the European level. It is important to note that data sharing could be performed at different levels of granularity. When data are shared at the single patient level, a key feature of an effective anonymization strategy is that it should be univocal, so that a patient receives the same anonymous ID each time his personal ID is being anonymized. This approach guarantees the possibility of tracking the same person through different data sources at the single country level (Figure 2).

At the same time, the anonymization technique must comply with the policies and constraints defined in national and international legislation and guidelines. Although many of these norms were developed in response to very different historical conditions, including technologies that have now been superseded, they have to be kept into account.<sup>35</sup> Furthermore, governance and ethical principles could have an important role in defining additional constraints on the type of approach used for anonymization. These in turn may influence if and how a database can be used in the study of multiple databases.

At the moment, there is no gold standard to perform multiple health care database integration among different countries and different health systems. One of the challenges in the fields of health services research is to promote the change from a fragmented to a harmonized approach defining sets of minimum data elements and agreed methods of and tools for harmonization and integration of data. In this sense, our study represented an effort to match EIP on AHA objectives of integration and harmonization in the field of large health-related databases.

By aggregating data from heterogeneous data sources and from large numbers of patients, the application of specific big data methodologies in the domain of clinical medicine and public health fulfils the goal of facilitating innovation, carrying out evidence-based research more efficiently and serving as the foundation for a more adaptive health care system.

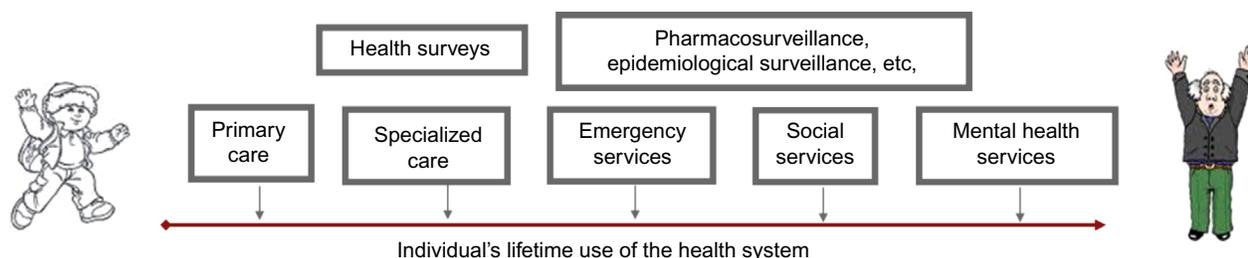
This article analyzed a limited number of databases, those available by the members of EIP on AHA Adherence Action Group. They are distributed on the EU territory and represent a cluster of data from three different countries (Italy, Spain, and the Republic of Ireland). However, the methodology used to obtain information and characterize these databases can be easily expanded to other interested stakeholders for future data integration. It is important to highlight that, at the current stage, this article aims to suggest a general framework to gather information necessary for health database integration, although some specificities may not be modeled by the current version of the survey employed. However, these features can be described by textual notes or by additional fields. This article did not treat legal and ethical aspects, since the focus of this work was on the technical and semantic feasibilities of interoperable databases from different countries and information at the individual level was out of the scope of this work.

## Future steps

However, the sharing processes pose many significant challenges such as ensuring interoperability both at a technical and semantic level. The first step to take in order to create such systems is the definition of a minimum dataset and to identify data gaps for sharing information at the European level. The next step will include the development of sharing models complying with ethical and legal constraints such as privacy policies and country-specific laws.

## Conclusion

Big data analytics and reasoning techniques could be used to deliver advanced health care services and to develop a



**Figure 2** Integration of diverse data sources from different levels of care.

systematic data gathering and integration at cross-national levels capable of providing support to both researchers and policy makers. In the framework of the EIP on AHA, our study represents the first step toward an approach for integration and harmonization of large health-related databases, with the final goal to create a structure capable of supporting multicountry health care research projects.

## Acknowledgments

The authors would like to thank the participants of the Collaborative Work on Adherence to Medical Plans of A1 Action Group of EIP on AHA. No funding was provided for the development of this study.

## Author contributions

All authors contributed toward data analysis, drafting, and critically revising the paper, gave final approval of the version to be published, and agreed to be accountable for all aspects of the work.

## Disclosure

The authors report no conflicts of interest in this work.

## References

- Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. *Annu Rev Public Health*. 2011;32:91–108.
- Harpe SE. Using secondary data sources for pharmacoepidemiology and outcomes research. *Pharmacotherapy*. 2009;29(2):138–153.
- Wettermark B. The intriguing future of pharmacoepidemiology. *Eur J Clin Pharmacol*. 2013;69(suppl 1):43–51.
- Hennessy S. Use of health care databases in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 2006;98(3):311–313.
- Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. Administrative record linkage as a tool for public health research.
- Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. *Yearb Med Inform*. 2014;9:14–20.
- Ayanian JZ. Using administrative data to assess health care outcomes. *Eur Heart J*. 1999;20(23):1689–1691.
- Schneeweiss S, Avorn J. A review of uses of healthcare utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323–337.
- Moriarty F, Bennett K, Fahey T, Kenny RA, Cahir C. Longitudinal prevalence of potentially inappropriate medicines and potential prescribing omissions in a cohort of community-dwelling older people. *Eur J Clin Pharmacol*. 2015;71(4):473–482.
- Onder G, Bonassi S, Abbatecola AM, et al; Geriatrics Working Group of the Italian Medicines Agency. High prevalence of poor quality drug prescribing in older individuals: a nationwide report from the Italian Medicines Agency (AIFA). *J Gerontol A Biol Sci Med Sci*. 2014;69(4):430–437.
- Iolascon G, Gimigliano F, Orlando V, Capaldo A, Di Somma C, Menditto E. Osteoporosis drugs in real-world clinical practice: an analysis of persistence. *Aging Clin Exp Res*. 2013;25(suppl 1):S137–S141.
- Casula M, Catapano AL, Piccinelli R, et al. Assessment and potential determinants of compliance and persistence to antiosteoporosis therapy in Italy. *Am J Manag Care*. 2014;20(5):e138–e145.
- Calderón-Larrañaga A, Gimeno-Feliu LA, González-Rubio F, et al. Polypharmacy patterns: unravelling systematic associations between prescribed medications. *PLoS One*. 2013;8(12):e84967.
- Calderón-Larrañaga A, Abad-Diez JM, Gimeno-Feliu LA, et al. Global health care use by patients with type-2 diabetes: does the type of comorbidity matter? *Eur J Intern Med*. 2015;26(3):203–210.
- Riera-Guardia N, Saltus CW, Bui CL, et al [webpage on the Internet]. Changes in the Landscape of Health Care Database Research From 2000 to 2011 (RTI Press Publication No. RR-0019-1308). Research Triangle Park, NC: RTI Press; 2013. Available from: <http://www.rti.org/rtipress>.
- European Commission [webpage on the Internet]. Innovation Union. Available from: [http://ec.europa.eu/research/innovation-union/index\\_en.cfm](http://ec.europa.eu/research/innovation-union/index_en.cfm). Accessed April 4, 2015.
- European Commission [webpage on the Internet]. Europe 2020. Available from: [http://ec.europa.eu/europe2020/index\\_en.htm](http://ec.europa.eu/europe2020/index_en.htm). Accessed April 4, 2015.
- European Commission [webpage on the Internet]. European Innovation Partnership on Active and Healthy Ageing. Available from: [http://ec.europa.eu/research/innovation-union/pdf/active-healthy-ageing/a1\\_action\\_plan.pdf#view=fit&pagemode=none](http://ec.europa.eu/research/innovation-union/pdf/active-healthy-ageing/a1_action_plan.pdf#view=fit&pagemode=none). Accessed April 4, 2015.
- Salisbury C, Johnson L, Purdy S, Valderas JM, Montgomery AA. Epidemiology and impact of multimorbidity in primary care: a retrospective cohort study. *Br J Gen Pract*. 2011;61(582):e12–e21.
- Healthcare Information Technology Standards Panel. Available from: <http://www.hitsp.org/>. Accessed April 4, 2015.
- HIMS. Federal Health Information Model. Available from: <http://www.fhims.org/>. Accessed April 4, 2015.
- BRIDG [homepage on the Internet]. The Biomedical Research Integrated Domain Group (BRIDG) Model. Available from: <http://www.bridgmodel.org/>. Accessed April 4, 2015.
- HL7 International [webpage on the Internet]. HL7 Reference Information Model (RIM). Available from: <http://www.hl7.org/implementation/standards/rim.cfm>. Accessed April 4, 2015.
- i2b2 [homepage on the Internet]. Informatics for Integrating Biology & the Bedside. Available from: <https://www.i2b2.org/>. Accessed April 4, 2015.
- Trifiro G, Coloma PM, Rijnbeek PR, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? (Review). *J Intern Med*. 2014;275:551–561.
- Blake KV, Trifiro G, Bourke A, et al [webpage on the Internet]. On Behalf of the ENCePP Working Group on Data Sources and Multi-Database Studies. Survey of Methodologies for European Union Publicly Funded Multi-Database Safety Studies. Available from: [http://www.encepp.eu/publications/documents/Survey\\_Multi-source\\_studies.pdf](http://www.encepp.eu/publications/documents/Survey_Multi-source_studies.pdf). Accessed May, 2015.
- EU-ADR web site. Available from: [http://synapse-pi.com/new\\_web/wp-content/uploads/2013/12/EU-ADR-project\\_flyer\\_20111.pdf](http://synapse-pi.com/new_web/wp-content/uploads/2013/12/EU-ADR-project_flyer_20111.pdf). Accessed April 4, 2015.
- PROJECT web site. Available from: <http://www.imi-protect.eu/>. Accessed April 4, 2015.
- Trifirò G, Patadia V, Schuemie MJ, et al; EU-ADR Group. EU-ADR healthcare database network vs. spontaneous reporting system database: preliminary comparison of signal detection. *Stud Health Technol Inform*. 2011;166:25–30.
- Oliveira JL, Lopes P, Nunes T, et al. The EU-ADR web platform: delivering advanced pharmacovigilance tools. *Pharmacoepidemiol Drug Saf*. 2013;22(5):459–467.
- Abbing-Karahagopian V, Kurz X, de Vries F, et al. Bridging differences in outcomes of pharmacoepidemiological studies: design and first results of the PROTECT project. *Curr Clin Pharmacol*. 2014;9(2):130–138.

32. SHARE web site [homepage on the Internet]. SHARE – Survey of Health, Ageing and Retirement in Europe. Available from: <http://www.share-project.org/>. Accessed April 4, 2015.
33. Malter F [homepage on the Internet]. Fieldwork Monitoring in the Survey of Health, Ageing and Retirement in Europe (SHARE). Survey Methods: Insights from the Field 2013. Available from: <http://survey-insights.org/?p=1974>. Accessed April 4, 2015.
34. Michelsen K, Brand H, Achterberg P, Wilkinson, J. *Promoting Better Integration of Health Information Systems: Best Practices and Challenges*. Copenhagen: WHO Regional Office for Europe; 2015. [Health Evidence Network (HEN) Synthesis Report].
35. Vayena E, Salathé M, Madoff LC, et al. Ethical challenges of big data in public health. *PLoS Comput Biol*. 2015;11(2):e1003904.

## Supplementary materials

Database identification and description	
1. Responsible organization	Please fill in with organization name
2. Owner	Please indicate the owner of the data, if it differs from your organization
3. Database name	Please fill in with database name
4. Database description	Please provide a short description of your database (use max 150 words)
5. Scope	Please fill in with the scope of the database as it relates to adherence to care plans (use max 300 words)
6. Data sources	Please provide information about which data sources are currently contributing to your database:
	<input type="checkbox"/> Patient personal data
	<input type="checkbox"/> Drug Prescriptions
	<input type="checkbox"/> Hospital discharge records
	<input type="checkbox"/> Medical examinations
	<input type="checkbox"/> Clinical information
	<input type="checkbox"/>
7. Geographic area	Please provide information about the geographic area covered by your database (use max 300 words)
8. Population covered:	
9. How many people are covered by your database	
10. Age span covered	
11. Whether the database is aimed to specific subset of population	
12. Data years available	Please indicate which years are available for your database
13. System category	<input type="checkbox"/> Major application: performs clearly defined functions for which there is a readily identifiable security consideration and need
	<input type="checkbox"/> General support system: provides general ADP or network support for a variety of users and applications
14. Operational status	<input type="checkbox"/> Operational <input type="checkbox"/> Under development <input type="checkbox"/> Undergoing a major modification
15. Database Management System	<input type="checkbox"/> Oracle <input type="checkbox"/> Microsoft SQL Server <input type="checkbox"/> PostgreSQL <input type="checkbox"/> MySQL <input type="checkbox"/> Microsoft Access <input type="checkbox"/> SQLite <input type="checkbox"/> Sybase <input type="checkbox"/> Teradata <input type="checkbox"/> Other (specify).....
16. Periodicity	Please provide information about how frequently your database is updated (use max 300 words)
17. Systems using the database	Please identify the systems that are currently using the database, if any (use max 300 words)
18. Relationship to other databases	Please indicate whether the database will supersede or interface with other databases, and specify which one(s) (use max 300 words)
19. Availability of technical specifications	Please state if technical documents are available for your database (use max 300 words)
20. Acronyms and abbreviations	Please provide a list of the acronyms and abbreviations used in this document and the meaning of each, if necessary (use max 300 words)
21. Other information of interest	Please indicate whether there are other information that should be known about your database (use max 300 words)

22. Relevant fields		Please fill in the following table indicating for each data source what are the relevant fields available in your database. Eg, hospital discharge records – Administrative-patient code, ICD-9 diagnosis code, patient's birthdate and sex, etc. PLEASE NOTE THAT DATA SOURCES IN THIS SECTION SHOULD MATCH THOSE INDICATED IN SECTION 6 (Data Sources)		
Data source		Relevant fields		
23. Points of organizational contact		Please provide a list of the points of organizational contact (POCs) that may be needed by the document user for informational and troubleshooting purposes. Include type of contact, contact name, department, telephone number, and email address (if applicable) Please add more rows if needed.		
Type of contact	Contact name	Department	Telephone number	E-mail address
<input type="checkbox"/> Helpdesk				
<input type="checkbox"/> Development				
<input type="checkbox"/> Maintenance				
<input type="checkbox"/> Operations				
<input type="checkbox"/> Privacy				
<input type="checkbox"/>				

**Figure S1** A template to collect information about database specifications

1. Responsible organization	Please fill in with organization name
2. Research group	Please fill in with research group name
3. Objective of the study	Please indicate the objectives of the study (use max 150 words)
4. Data source	Please provide information about which data sources are currently contributing to your study (more than one choice is allowed):
	<input type="checkbox"/> Patient personal data
	<input type="checkbox"/> Drug prescriptions
	<input type="checkbox"/> Hospital discharge records
	<input type="checkbox"/> Medical examinations
	<input type="checkbox"/> Clinical information
	<input type="checkbox"/>
5. Study population	Please provide key elements of the study. Give the eligibility criteria, and the sources and methods of selection of participants, describe the setting (use max 300 words)
6. Time frame	Please provide relevant dates, including periods of recruitment, exposure, follow-up, and data collection.
7. Type of medication	Please indicate type of medication included in the study (use ATC code)
8. Adherence indicator	Please provide information about adherence indicators (more than one choice is allowed):
	<input type="checkbox"/> Proportion of days covered (PDC)
	<input type="checkbox"/> Medication possession ratio (MPR)
	<input type="checkbox"/> Length of stay in treatment (persistence)
	<input type="checkbox"/> Presence/absence
	<input type="checkbox"/> Refill
	<input type="checkbox"/> Self-reported adherence
<input type="checkbox"/>	
<input type="checkbox"/>	

9. Other study variables	Please define all variables of interest in the analyses (use max 150 words)
10. Studied outcomes	Please clearly define all outcomes (use max 150 words)
11. Available results	Please summarize key results (use max 300 words)
12. Other observations	Please provide other observations, if any (use max 150 words)

**Figure S2** A template to collect information about observational studies on the measurement of adherence in the older population based on health-related databases

### ClinicoEconomics and Outcomes Research

Dovepress

#### Publish your work in this journal

ClinicoEconomics and Outcomes Research is an international, peer-reviewed open-access journal focusing on health technology assessment, pharmacoeconomics and outcomes research in the areas of diagnosis, medical devices, and clinical, surgical and pharmacological intervention. The economic impact of health policy and health systems

organization also constitute important areas of coverage. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/clinicoeconomics-and-outcomes-research-journal>