Towards Health Information Integration

Walter Gall[1], Wolfgang Dorda[1], Georg Duftschmid[1] , Gottfried Endel[2], Karl Fröschl[3], Wilfried Grossmann[3], Milan Hronsky[1]

[1] *Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna*
[2] *Main Association of the Austrian Social Security Institutions*
[3] *Institute for Scientific Computing, University of Vienna*

Abstract

Background: While data based on health statistics (S-data) provide a summarized view of the health of a general population, data based on electronic health records (EHR-data) provide information about individual patients. Serving fairly different purposes, the two approaches to health information have evolved more or less independently. However, various benefits of using EHR-data in connection with public health issues have been identified and discussed.
Objectives: The conceptual differences between these two approaches and the potential benefits of integration are discussed. A schematic illustration of the integration of EHR-data and S-data is outlined to analyze an integration scenario.
Methods: As a test case we used reimbursement data of the Main Association of Austrian Social Security Institutions (EHR-data) and Austrian data of the European Community Health Survey (S-data). The time frame considered was restricted to the period from 2006 to 2007, and the prevalence of diabetes mellitus in Austria was selected as the exemplary subject of interest.
Results: With respect to specific medical concepts, comparisons between EHR-data and S-data are clearly feasible. EHR-data are potentially valid substitutes of S-data and can provide detailed evidence for health reporting. For diabetes the difference of the prevalence between EHR-data and S-Data was only 1% for whole Austria.
Conclusions: The pilot study yielded encouraging results. With respect to specific medical concepts, comparisons between EHR-data and S-data are clearly feasible. EHR-data are potentially valid to substitute or supplement surveys and can provide detailed evidence for health reporting.
Keywords: Epidemiological research, public health, electronic health records, data analysis

Background

A number of initiatives in the European Union have been focused on the development of eHealth systems [1]. The same is true for Austria [2]. These systems will serve as new and powerful sources of health information about the residents of a

country. The electronic health record (EHR) is the informational basis of such eHealth systems [3]. Flexible exchange of data between different distributed EHR systems as well as secure and confidential storage of patient records to facilitate communication between health professionals about the status of an individual, have been given significant attention. Hence, this type of information is characterized by a medical view onto **individual patients**. In the present report we use the term **EHR-data** to refer to patient-related health information in a rather wide sense.
A further established source of data concerning health, based on health statistics, provides a **summarizing view** onto the health of a specific population. Such health statistics are usually based on surveys or registers. In the following, such information is referred to as **S-data.**
Serving rather diverse purposes, the two approaches to health information have evolved more or less independently. However, while secondary use of EHR-data and related administrative data is not straightforward [4][5], these nevertheless provide valuable information about the health status of a population and thus support health policy and research in the field of health services. Interoperable EHR models and reliable data protection methods might additionally improve secondary data usage. Various benefits of using EHR-data in connection with public health issues have been identified and discussed [6][7][8]:

- **Augmentation of S-data by EHR-data.** S-data provide information based on the use of a rather general medical terminology and omit details (such as diabetes, with no specific information being provided about the type of diabetes or the method of treatment). Such details may be found in EHR-data and would provide a more detailed picture of the health status of an entire population.
- **Reduction of response burden.** In some cases the desired information about the health status of a group can be inferred from EHR-data (e.g. medical and socio-demographic information in cases of cancer). This may reduce the number of questions presented to the surveyed persons.
- **Comparison between subjective and objective wellbeing.** In view of the strongly subjective aspect of all health issues, EHR-data and S-data complement each other. EHR-data are focused to a greater extent on objective health criteria while S-data based on surveys concern the individual's personal well-being. A comparison of the two data sources may enhance our understanding of the discrepancies between the two views.
- **Medical research.** To evaluate the implications of data obtained from medical research for the purpose of formulating health policies, the analyst must combine data from clinical trials with those about the health status of a population.

Objective

In this study we present a few basic concepts to bridge the gap between EHR-data and S-data. We discuss the principal differences between the two approaches towards health information and present a conceptual model of health data integration. As medical test cases we use data about diabetes and chronic pain. Data obtained from the Austrian part of a European health survey [9] serve as S-data while administrative (i.e., reimbursement) data of the Main Association of Austrian Social Security Institutions (HV) are used as proxy for EHR-data proper. As a basis for further analysis we will determine whether the prevalence of diabetes and chronic

pain of the patient cohorts derived from EHR-data, can be compared with the figures obtained from the health survey. In the present study we report the preliminary results of the sub-project focused on diabetes.

Methods

1. Conceptual differences between EHR-data and S-data

Coherent joint use of EHR-data and S-data requires clear expression of the conceptual differences between the two different sources of information. The major dimensions of conceptual divergence typically refer to:

(a) **Underlying population and representativeness.** Contrary to the medical view which is mainly focused on individual patients, the statistical approach views the entire population of a country or a subgroup of interest, such as individuals with specific symptoms. In general EHR-data encompass a subset of the entire population which is not selected according to its representativeness in the statistical sense, but according to the occurrence of medical interventions. Depending on the scope of EHR-data, a complete survey may be available for certain types of data (e.g. administrative data) or medical circumstances (e.g. notifiable diagnoses). A partial survey, on the other hand, always comprises just a random sample.

(b) **Level of detail.** Usually the attributes of EHR-data provide much more detailed information about an individual's health status in respect of clinical parameters (e.g. detailed information about the use of medical services or medication of a person as a time series). Such detailed information is rarely found in S-data. Additionally, EHR-data can provide a longitudinal picture of the interventions performed on a patient, whereas S-data would refer to the patient's health status at a specific time point.

(c) **Objective versus subjective information.** EHR-data are usually obtained from medical interventions and recorded by health professionals. In contrast to such objective information, S-data are usually reported by individuals and provide a subjective statement about the surveyed person's health status. Such statements may differ from a medical diagnosis, particularly when the investigator is interested in obtaining information about a person's well-being, which can hardly be captured by objective criteria.

(d) **Data structure.** Like most statistical data, the structure of S-data is rather simple and normalized, and can be characterized either by a case-by-variates matrix of the surveyed persons, together with weights for generalization or by a table of counts at the population level. The structure of EHR-data may be complex. In addition to records of measurements, EHR-data may include semi-structured data such as written reports or multi-media data files.

2. Schematic integration of EHR-data and S-data

As EHR-data generally provide information at a personal level while S-data constitute information only at anonymized level, an integration scenario is only possible in aggregated form at a population level.

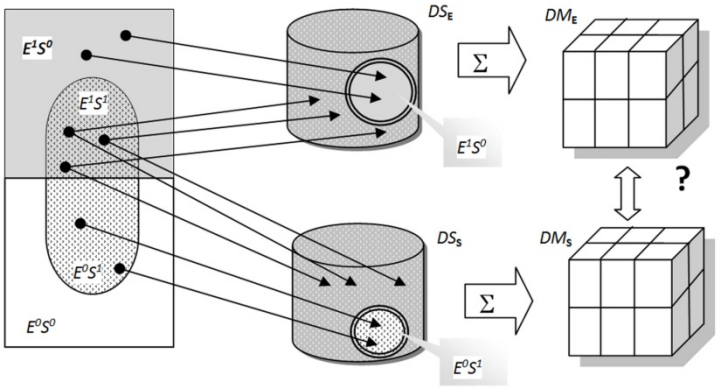

Figure 1: Schematic integration of EHR-data and S-data

Figure 1 is a schematic illustration of such integration. The entire population (rectangle) is split into two parts, namely E1 and E0 (squares with a gray and a white background). E1 are those persons for whom EHR-data are available while E0 consists of those for whom no EHR-data exist. S1 represent those persons for whom S-data is available (patterned oval object) while S0 represents the remainder of the population for whom no S-data are available. S-data are collected in the data store named DSS and subsequently aggregated in a data mart named DMS. EHR-data are collected in a data store named DSE and then aggregated in a data mart named DME. Due to the rather complex and diverse structures of EHR-data, it is usually difficult to specify those EHR-data which should constitute the DME data mart. Even in cases of identical dimensions for both data marts and identical summary attributes, the two cubes are not exactly comparable because of the two sets E1S0 and E0S1, which represent those persons for whom EHR-data, but no S-data (E1S0) exist, and those persons who have been surveyed but lack EHR-data (E0S1). In case of weighted survey information one can resolve these differences using the weights, in other cases the calculation may be rather cumbersome.

Results

As a practical case of the integration concept we analyzed a set of existing Austrian EHR-data and S-data sources with regard to their comparability of clinical concepts and dimensions. Eventually, for the purpose of methodological comparison, reimbursement data of the Main Association of Austrian Social Security Institutions (EHR-data) and Austrian data of the European Community Health Survey [9] were considered the most suitable. The time frame was confined to the period from 1.1.2006 to 31.12.2007, and the prevalence of diabetes mellitus and chronic pain in Austria was selected as the exemplary subject of interest. In the following the results of the diabetes project are described.

The chosen EHR-data did not contain encoded diagnoses but information about prescriptions of pharmaceuticals for diabetes (ATC codes A10A, A10B or A10X). In the juxtaposed health survey (S-data), five questions refer to the diagnosis "diabetes", one of which inquiries into the intake of pharmaceuticals for diabetes.

Based on these data sources, the data marts DME and DMS were defined using the dimensions age, gender, and geographic region, which were available in both of the datasets. S-data comprise the summary attributes “occurrence of diabetes”, “medical

diagnosis of diabetes", and "treatment of diabetes with pharmaceuticals", together with weights for each case. For EHR-data, different subversions of the summary attribute "treatment of diabetes with pharmaceuticals" were aggregated on the basis of the reported frequency of prescriptions within the selected time period.

Comparison of the two resulting data marts with respect to the four conceptual dimensions outlined in Section 0 produced the following results:

(a) **Underlying population and representativeness.** S-data only consider residents in Austria older than 15 years of age (approximately 7 million persons), whereas EHR-data contain all persons registered at the social insurance, independent of age and permanency of residence. In this case the EHR-data provide a more accurate picture because, in contrast to the random sample (S-data), a complete survey is performed here. Alignment regarding age was no problem. With respect to the place of residence, the two populations showed comparable figures for the summary attribute "treatment of diabetes with pharmaceuticals". Table I shows a comparison of the populations older than 15 years of age according to provinces and gender. EHR-data contain 350,041 persons with diabetes mellitus while S-data comprise 353,039 persons. For EHR-data, this signifies an under-registration of 11% in women, an over-registration of 9% in men, and an overall under-registration of 1%. The reasons for the differences may be attributed to different subjective evaluation of diseases by women and men and to regional differences.

Table I: Comparison of persons with diabetes mellitus older than 15 years of age in the populations of EHR-data and S-data, divided according to political provinces and gender.

	males		females		sum	
province	EHR-data	S-data	EHR-data	S-data	EHR-data	S-data
Burgenland	7.153	7.233	8.111	6.563	15.264	13.796
Carinthia	10.046	9.443	10.482	13.624	20.528	23.067
Lower Austria	36.907	26.235	37.347	39.333	74.254	65.568
Upper Austria	26.632	16.803	25.667	30.821	52.299	47.624
Salzburg	8.828	7.925	8.525	7.930	17.353	15.855
Styria	23.577	25.381	25.734	27.608	49.311	52.989
Tyrol	9.984	11.748	10.428	12.098	20.412	23.846
Vorarlberg	5.628	4.886	5.627	6.488	11.255	11.374
Vienna	44.836	48.396	43.814	50.524	88.650	98.920
missing	314		401		715	
sum	173.905	158.050	176.136	194.989	350.041	353.039
difference		- 9%		11%		1%

(b) **Level of detail.** Due to the fact that medical prescriptions distinguish between pharmaceuticals for insulin-dependent and non-insulin-dependent diabetes, EHR-data give a more detailed picture of the disease diabetes mellitus than the S-data. Additionally, due to the sample sizes, prevalence could not be analyzed on a fine district level within S-data, but only on a higher granularity on a province level. Within EHR-data these analyses on a fine spatial granularity are possible. As an example, figure 2 shows the prevalence of diabetes mellitus derived from EHR-data for females and males in the age of 45 to 59 years on district level (NUTS 3).

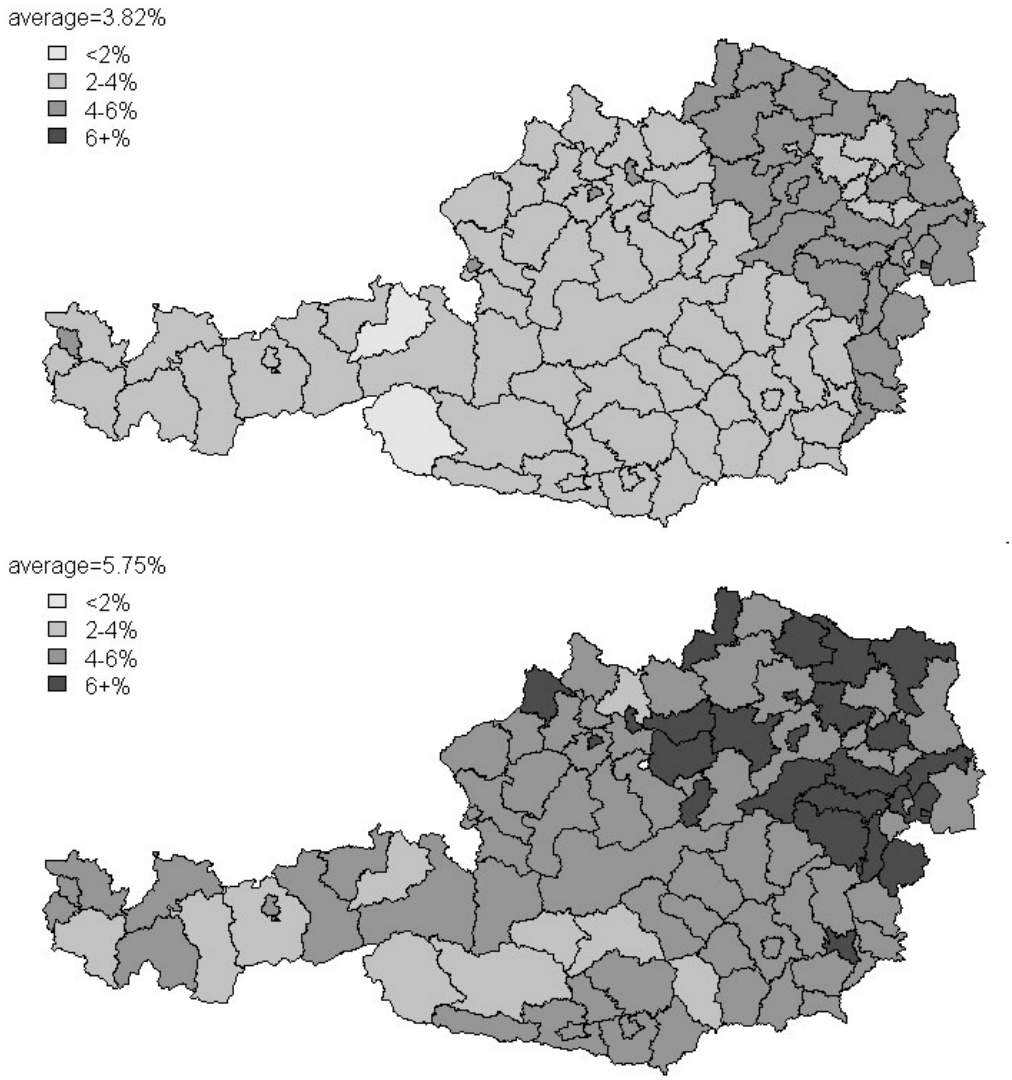

Figure 2: Prevalence of diabetes mellitus in Austria of females (above) and males (below) in the age of 45 to 59 years derived from EHR-data

(c) **Objective versus subjective information.** The concepts used in the two datasets seem to be rather identical in this case. When the surveyed persons answered the question "Did you ever have diabetes" in the affirmative, the next question "Did you take medication for this purpose?" yields subjective data which, however, can be compared well with the objective data concerning dispensed medication.

(d) **Data structure.** The EHR-data of the Main Association are billing data derived from the various health insurance companies in the various provinces. Due to missing or contradictory data, certain interrelationships between the data had to be established by means of statistical matching procedures.

Discussion

Generally speaking, the pilot study yielded encouraging results. With respect to specific medical concepts, comparisons between EHR-data and S-data are clearly feasible. EHR-data are potentially valid substitutes of S-data and can provide detailed evidence for health reporting. However, it might be possible to improve several aspects by way of data quality (e.g. different levels of quality in the provinces) as well as constitution of the populations (tourists, externally insured persons) and thus enhance the quality of the results.

Secondary use of EHR-data requires, in addition to the collection of health information on a personal level, careful administration of data. Problems of privacy can be avoided effectively by processing individual EHR-data on aggregated levels only. Further alternatives would be pseudonymization or k-anonymity.

Conclusion

Integrating health data from different sources may help to improve the efficiency as well as coverage of current health statistics. However, a thorough comprehension of the derivation of data and processes is required for this purpose. While the supply of data could be supported by EHR-based generic retrieval models [10] and query languages [11] (which are currently being developed), a legal framework regulating secondary use of EHR-data is required. In keeping with other countries (like the USA [12]), basic legal, technical and organizational preconditions will be established in Austria to expedite such eHealth evaluations.

Acknowledgement:

The project received financial support from the Main Association of Austrian Social Security Institutions.

Clinical Relevance Statement:

Integration of objective and subjective information enables medical specialists to round up their knowledge about a specific disease. Evaluation of EHR-data will save costs by partially replacing expensive surveys or supplementing the information obtained from such surveys.

References

[1] Europe's Information Society - Quality of Life: health. [cited 2011 April 20]. Available from: http://ec.europa.eu/information_society/tl/qualif/health/index_en.htm.

[2] Dorda W, Duftschmid G, Gerhold L, Gall W, Gambal J. Austria's path toward nationwide electronic health records. Methods Inf Med. 2008; 47(2):117-23.

[3] Blobel B. Advanced EHR architectures--promises or reality. Methods Inf Med. 2006; 45(1):95-101.

[4] Cruz-Correia R, Rodrigues P, Freitas A, Almeida F, Chen R, A. C-P. Data Quality and Integration Issues in Electronic Health Records. In: Hristidis V, editor. Information discovery on electronic health records. London: CRC Press; 2010. p. 55-95.

[5] Gall W, Grossmann W, Duftschmid D, Wrba T, Dorda D. Analyses of EHRs for research, quality management and health politics, Stud Health Technol Inform. 2008; 136:425-430.

[6] Bain M, Chalmers J, Brewster D, Routinely collected data in national and regional databases – an under-used resource. J of Public Health Medicine. 1997; 19(4):413-418.

[7] Kukafka R. Ancker J, et al., Redesigning electronic health record systems to support public health. J of Biomedical Informatics. 2007; 40:398-409.

[8] De Clerq E, Van Casteran V, et al., Research networks: can we use data from GP's Electronic Health Records? Stud Health Technol Inform. 2006; 124:181-186.

[9] Statistik Austria. Healthstatistics. [cited 2011 April 20]. Available from: http://www.statistik.at/web_de/dynamic/statistiken/gesundheit/publdetail?id=4&listid=4&detail=457.

[10] Austin T, Kalra D, Tapuria A, Lea N, Ingram D. Implementation of a query interface for a generic record server. Int J Med Inform. 2008; 77(11):754-64.

[11] Ma C, Frankel H, Beale T, Heard S. EHR Query Language (EQL) – A Query Language for Archetype-Based Health Records. Stud Health Techn Inform. 2007; 127:397-401.

[12] Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an AMIA White Paper. J Am Med Inform Assoc. 2007; 14(1):1-9.