Appl. Math. Inf. Sci. **8**, No. 1L, 1-6 (2014)

Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/081L01

Quality in Hospital Administrative Databases

Alberto Freitas[1,2,*], *Juliano Gaspar*[1,2], *Nuno Rocha*[1,2], *Goreti Marreiros*[3] *and Altamiro da Costa-Pereira*[1,2]

[1] Department of Health Information and Decision Sciences, Faculty of Medicine, University of Porto, Portugal
[2] CINTESIS - Center for Research in Health Technologies and Information Systems, Portugal
[3] GECAD - Knowledge Engineering and Decision Support Group, Institute of Engineering, Polytechnic of Porto, Portugal

Received: 7 Apr. 2013, Revised: 2 Aug. 2013, Accepted: 3 Aug. 2013
Published online: 1 Apr. 2014

Abstract: The clinical content of administrative databases includes, among others, patient demographic characteristics, and codes for diagnoses and procedures. The data in these databases is standardized, clearly defined, readily available, less expensive than collected by other means, and normally covers hospitalizations in entire geographic areas. Although with some limitations, this data is often used to evaluate the quality of healthcare. Under these circumstances, the quality of the data, for instance, errors, or it completeness, is of central importance and should never be ignored. Both the minimization of data quality problems and a deep knowledge about this data (e.g., how to select a patient group) are important for users in order to trust and to correctly interpret results. In this paper we present, discuss and give some recommendations for some problems found in these administrative databases. We also present a simple tool that can be used to screen the quality of data through the use of domain specific data quality indicators. These indicators can significantly contribute to better data, to give steps towards a continuous increase of data quality and, certainly, to better informed decision-making.

Keywords: administrative data, hospital information systems, data quality, data quality problems, measuring performance, business intelligence, international classification of diseases

1. Introduction

This paper aims to discuss the use and the quality of administrative data, specifically data from hospital discharges, the originally called minimum basic data set (MBDS) [1]. The concept of MBDS, firstly formalized in 1973 [2], has been defined as the core of patient information with the most commonly available items and the most extensive range of usages.

Administrative data (also known as administrative databases, MBDS, or secondary data) are data sets that have been created in the area of health services usually for billing purposes [3].

The clinical content of administrative databases includes, among others, patient demographic characteristics and codes for diagnoses and procedures. This data is standardized, clearly defined, readily available, less expensive, and normally covers the majority of the hospitalizations. Although with some limitations, this data is often used to construct measures and evaluate the quality of healthcare [4].

The increased availability of administrative data induces an increase in the number of research studies using this electronic secondary data. Nevertheless, there are some special considerations when using secondary data for research. In this situation, special attention should be given to ensure that, for a specific study, the relevant data is selected, the correct selection of codes is well defined (e.g., inclusion and exclusion of ICD-9-CM[1] codes for a specific diagnosis) [5]. In this context, it is important to have a good understanding of coding and classification systems, to be comfortable with the usual coding protocols for some diseases and be aware of changes in coding rules through the years [6,7]. Practitioners familiarized with the special issues in this data can use it in research studies and can contribute to the improvement of healthcare.

This data is often used, for instance, to analyze trends in specific surgeries or medical conditions [8,9,10], to pharmacoepidemiology studies [5], in the study of comorbidities [11], and to developed quality indicators [12].

Administrative data usually includes basic patient characteristics, such as gender, age, residence, admission

[1] International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) - http://www.cdc.gov/nchs/icd/icd9cm.htm

* Corresponding author e-mail: alberto@med.up.pt

and discharge data, diagnoses (principal and secondary), procedures, length-of-stay, type of admission, discharge disposition of the patient, and expected source of payment. Other variables are also available, such as the Diagnosis Related Group (DRG) [1], created originally to identify the output of hospitals in a consistent, systematic and exhaustive manner.

The rest of this paper is organized as follows: Section 2 briefly explains a taxonomy for the classification of data quality problems (DQP); In Section 3 we give some examples and discuss possible implications of data quality problems in hospital administrative data; In Section 4 we present a tool for the production of performance indicators that includes a module for the detection and comparison of data quality problems; and, in Section 5, we present the concluding remarks.

2. Data Quality

The quality of data is a critical issue in all areas and, in particular, in the healthcare arena. Clinically, data quality is important for a correct diagnosis and treatment, but it is also essential to support health management and epidemiological research. Data quality is often defined as ”fitness for use”, that is, if the data is fit for a particular use or not [13]. Figure 1 concisely represents the information flow, from patient to administrative databases and data usage. In any of the represented steps many possible events or problems can contribute to the final quality of the administrative data.

Several authors, with little agreement, have described different dimensions for data to have quality. ISO/IEC 25012:2008 defines a general data quality model and presents fifteen data quality characteristics for Information Systems: accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability, and recoverability.

In a database perspective, i.e., considering only the quality of the data values or instances, data quality problems can be classified as (Figure 2) [14]: Missing Values [MV], Syntax Violation [SV], Domain Violation [DV], Incorrect Value [IV], Violation of Business Rule [VBR], Uniqueness Violation [UV], Existence of Synonyms [ES], Violation of Functional Dependency [VFD], Duplicate Tuples [DT], Approximate Duplicate Tuples [ADT], Inconsistent Duplicate Tuples [IDT], Referential Integrity Violation [RIV], Incorrect Reference [IR], Heterogeneity of Syntaxes [HS], Heterogeneity of Measure Units [HMU], Heterogeneity of Representation [HR], and Existence of Homonyms [EH]. These DQP can be found through the difference database layers (data source, relation, tuple, column, row, attribute). This taxonomy was used in the development of a data quality tool, briefly presented in section 4.

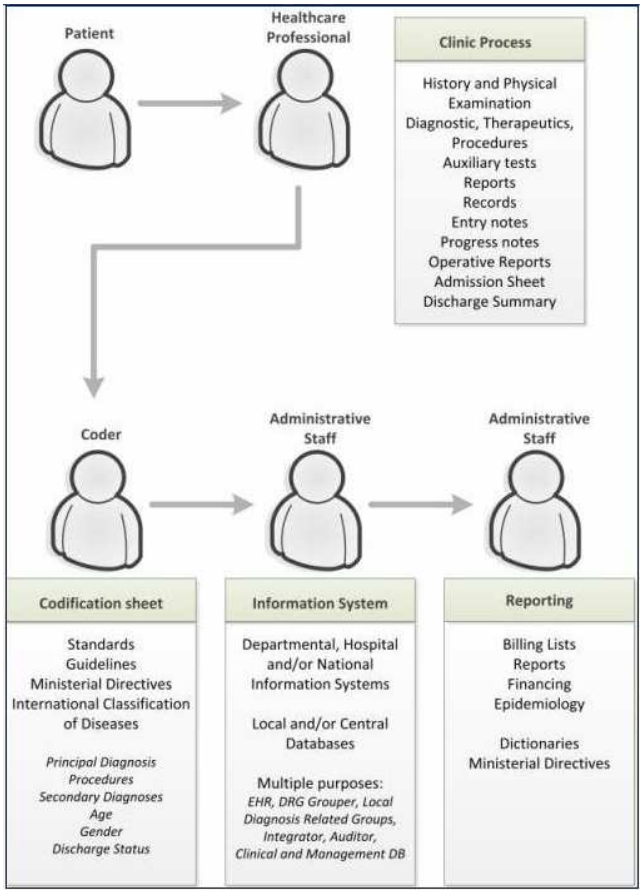

Figure 1: Information flow, from patient to administrative databases and data usage.

3. Examples of Data Quality Problems

In this section we present and discuss some examples of data quality problems found in administrative data. We used a national hospital database, with data from 10,586,118 inpatient episodes discharges, between 2000 and 2010, in public acute care hospitals of the National Health Service (NHS), representing about 85% of all inpatient stays in Portugal. The access to the data was provided by ACSS, I. P. (Administração Central do Sistema de Saúde, I. P.), the Ministry of Health’s Central Administration for the Health System.

3.1. Number of secondary diagnoses

Comorbidities are associated with health outcomes (increased in-hospital mortality and length of stay), and health care costs. Comorbidity assessment is important not only for health services research, epidemiological and clinical studies, but also for financing, health policy and health care planning.

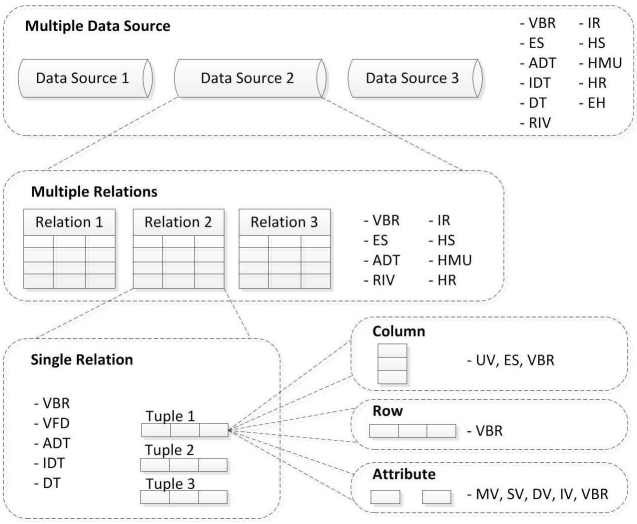

Figure 2: A data quality taxonomy, adapted from [14].

When using administrative data, comorbidities (pre-existing conditions) should always be controlled [15]. In these databases, comorbidities are identified through the ICD-9-CM codes of secondary diagnoses. In this context, the evolution in the quality and in the quantity of coded secondary diagnoses can influence the proportion of identified comorbidities and so, in any temporal analysis, these possible limitations should be considered and discussed.

As we can see in Figure 3, the number of coded secondary diagnoses and consequently the number of comorbidities are continuously increasing over years. This evolution is not related with an increase in the severity of treated patients but rather with the number of coded secondary diagnoses. As in other comorbidity studies, for instance the study by Elixhauser et al. [15], we excluded pediatric (age below 18 years) and obstetrical episodes.

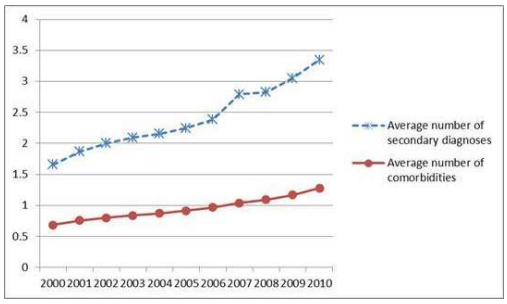

Figure 3: Evolution in the average number of coded secondary diagnoses and of extracted comorbidities, per hospital inpatient episode (for age 18 years and older and excluding obstetric admissions).

3.2. Ischemic Stroke

The evolution of ischemic stroke codification is a clear example of a situation where the lack of information about changes in the coding protocol can lead to misleading interpretations. Ischemic stroke (a stroke caused by thrombosis or embolism) does not have a direct entry in the ICD-9-CM index and, in Portugal, it was originally often classified, erroneously, with ICD-9-CM code 437.1. After 2004, Portuguese medical coders started to correctly code this situation with 434.91.

As we can see in Figure 4, the generalization of the correct coding practice took a period of 4 to 5 years. This long period for changes to be effective occurred, probably, because messages took time to reach all medical coders and also because the used ICD-9-CM versions are often not updated and are not uniform.

These changes in coding practices, if not known and considered, could clearly lead to erroneous conclusions in any clinical or epidemiological study of ischemic stroke.

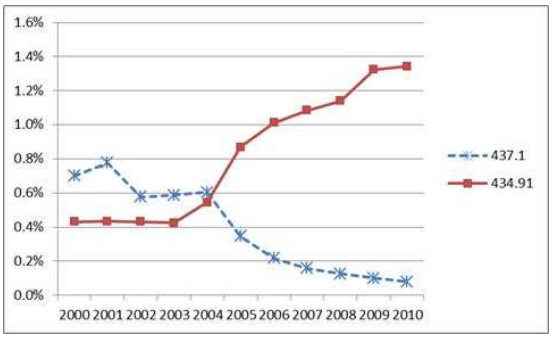

Figure 4: Evolution of Ischemic Stroke codification.

3.3. Birth Weight

Birth weight is a proxy of prematurity and therefore can be used to select high risk groups of newborns and also in risk adjustment. As in the work of Schwartz et al. [3], we also studied the percent of agreement between birth weight and the ICD-9-CM 5th digit coding. Our results also show that researchers can feel confident about the quality of this administrative variable in the identification of very low birth weight infants (table 1). However, for the range of 2000-2499 grams our results are significantly lower than the ones of Schwartz study (61.3% agreement, 23.5% less). This percentage is low and should be considered is studies that use birth weight (for instance, if used as a proxy for performance measurement, in the selection criteria).

Another potential implication of an incorrect value for birth weight is related to hospital budget. Birth weight is one of the variables used to group episodes into Diagnosis Related Groups (DRGs) and a small difference in birth

weight can originate a different DRG. For instance, in AP-DRG (All Patient DRGs) version 21, according to the Ministry of Health, DRG 622[2], for birth weight above 2499 grams, is priced 19 628 euros while DRG 615, for birth weight between 2000 and 2499 grams, is priced 31 298 euros. An incorrect birth weight can erroneously lead to the attribution of a different DRG, with direct implications in the hospital budget.

4. A Tool for the Detection of Data Quality Problems

We defined and implemented a module for the detection of data quality problems (DQP) and inconsistencies in inpatient and ambulatory episodes in administrative hospital databases. A diagram generically representing the implementation strategy is present in Figure 5. This data quality module identifies errors in common variables in health databases, and also includes domain-specific rules for the detection of errors in data from specific medical specialties.

This model detects and summarizes information for each DQP (type and number of errors) and allows the comparison of them over time and by hospital or type/group of hospitals. Implemented DQP rules validate whether the information contained in a record is consistent, meaning that it does not present incongruences between the values of a variable in relation to values of other variables in the same episode, and whether values are accordingly to regulations and coding guidelines used in Portugal (e.g. ICD-9-CM). These implemented DQP rules are classified according to the taxonomy presented in Section 2, as shown in table 2.

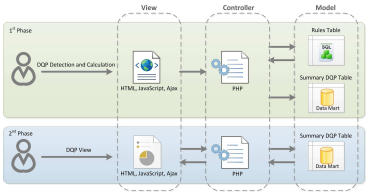

Figure 5: Model-View-Controller (MVC).

This module allows the user to edit the rules for the DQP detection. It also allows the user to create new rules for the detection of specific DQP, best suited to the user's context. For example, in the case of coding for Breast Cancer, for female gender, the used code should be 174.x (if DIAG="174%" then SEX="F"), while for male gender, the correct code is 175.x (if DIAG="175%" then SEX="M").

[2] DRG 615: Neonate, birth weight 2000-2499g, with significant operating room procedure, with multiples major problems; DRG 622: Neonate, birth weight >2499g, with significant operating room procedure, with multiples major problems.

Table 2: Number of rules grouped using the DQP taxonomy.

Abbrev.	Description	Nbr. of rules
MV	Missing values	28
SV	Syntax violation	14
DV	Domain violation	23
IV	Incorrect value	3
VBR	Violation of business rule	14
UV	Uniqueness violation	10
DT	Duplicate tuples	1
IR	Incorrect reference	1

With this tool it is possible to create reports for the detection of over 90 different types of DQP and related incongruences. Results are graphically simple and allow a direct reading of the most relevant values being an asset to help managing and controlling the efficiency of health facilities. The visualization of DQP, in a graphical format with data over years, is allowing users to, with only a visual analysis of the evolution of the problem, quickly identify possible relationships between the increase or decrease of a problem and for instance the date of introduction of a certain software application in the hospital, or the adoption of a new standard for procedures.

Figure 6 shows an example of the use of this DQ module where we can note a clear decrease in the number of DQP found in the variable 'birth weight'. The observed gradual decrease can be associated with data validation rules, performed during data entry, by the different electronic applications used on NHS hospitals during the last years.

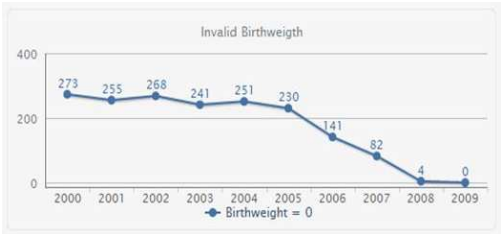

Figure 6: Evolution of DQP found in the variable 'birth weight'.

This module represents an important contribute for the detection of DQP in administrative and other health databases and, in addition, is an incentive to the improvement of the architecture of existing information systems. In fact, this tool also intends to alert for the impact of these errors in epidemiological studies, management evaluations, and health related policies.

5. Conclusion

In this article we explore issues related to quality in administrative databases. We present some data quality prob-

Table 1: Agreement between birth weight and ICD-9-CM code (for discharges in 2009 and 2010).

Actual Birth Weight	Birth weight code from DX 764.xx, 765.0x and 765.1x						
	500-749g	750-999g	1000-1249g	1250-1499g	1500-1749g	1750-1999g	2000-2499g
500-749g	90.52%						
750-999g		92.73%					
1000-1249g			85.48%				
1250-1499g				86.14%			
1500-1749g					85.54%		
1750-1999g						85.52%	
2000-2499g							61.27%
Difference for [3]	-0.67%	0.67%	-5.59%	-6.38%	-5.55%	-6.11%	-23.53%

lems and discuss their implications. Data producers and data consumers should be aware of the strengths and weaknesses of these databases. The completeness and accuracy in this data can only be improved at the source, and so hospitals should give attention and be committed in auditing and reporting the quality of both data and processes. The increase use of administrative data, the use of data quality and performance tools, and the implications in hospitals financing is, nowadays, a major incentive for hospitals to improve quality of care and decrease costs. The public available information about hospitals performance will also be a motivation for their continuous demand for quality.

Administrative data offers a useful and cost-effective source of information to monitor and evaluate the impact of healthcare policies at a local, regional or national level. Nevertheless much work is still needed to develop health information systems and increase the quality of this routinely collected data.

Acknowledgement

The authors wish to thank ACSS, for providing access to the data, and the support given by the research project HR-QoD - Quality of data (outliers, inconsistencies and errors) in hospital inpatient databases: methods and implications for data modeling, cleansing and analysis (project PTDC/SAU - ESA /75660/ 2006). This work is partially supported by FEDER Funds through the ”Programa Operacional Factores de Competitividade - COMPETE” program and by National Funds through FCT ”Fundação para a Ciência e a Tecnologia” under the project: FCOMP-01-0124-FEDER-PEst-OE/EEI/UI0760/2011.

References

[1] M. Casas and M.M. Wiley. *Diagnosis related groups in Europe: uses and perspectives*. Springer-Verlag Berlin, 1993.

[2] F.H. Roger, European Communities Commission. Directorate-General for Information Market, Innovation. Committee for Scientific, Technical Information, and Documentation. Biomedical Working Group. *The Minimum Basic Data Set for Hospital Statistics in the EEC: Review of Availability and Comparability*. Commission of the European Communities, 1981.

[3] R. M. Schwartz, D. E. Gagnon, J. H. Muri, Q. R. Zhao, and R. Kellogg. Administrative data for quality improvement. *Pediatrics*, 103(1 Suppl E):291–301, 1999.

[4] L. I. Iezzoni. Assessing quality using administrative data. *Ann Intern Med*, 127(8 Pt 2):666–74, 1997.

[5] S. E. Harpe. Using secondary data sources for pharmacoepidemiology and outcomes research. *Pharmacotherapy*, 29(2):138–53, 2009.

[6] R.J. Cruz-Correia, P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira. Data quality and integration issues in electronic health records. In Vagelis Hristidis, editor, *Information Discovery on Electronic Health Records*, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC, 2009. doi:10.1201/9781420090413-c4.

[7] A. Freitas, T. Silva-Costa, B. Marques, and A. Costa-Pereira. Implications of data quality problems within hospital administrative databases. In Panagiotis D. Bamidis and Nicolas Pallikarakis, editors, *IFMBE Proceedings*, volume 29 of *IFMBE Proceedings*, pages 823–826. Springer Berlin Heidelberg, 2010.

[8] J. F. Finks, N. H. Osborne, and J. D. Birkmeyer. Trends in hospital volume and operative mortality for high-risk surgery. *N Engl J Med*, 364(22):2128–37, 2011.

[9] J. D. Birkmeyer, A. E. Siewers, E. V. Finlayson, T. A. Stukel, F. L. Lucas, I. Batista, H. G. Welch, and D. E. Wennberg. Hospital volume and surgical mortality in the united states. *N Engl J Med*, 346(15):1128–37, 2002.

[10] A. J. Walkey, R. S. Wiener, J. M. Ghobrial, L. H. Curtis, and E. J. Benjamin. Incident stroke and mortality associated with new-onset atrial fibrillation in patients hospitalized with severe sepsis. *JAMA*, 306(20):2248–54, 2011.

[11] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J. C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali. Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. *Med Care*, 43(11):1130–9, 2005.

[12] G. H. Utter, A. M. Borzecki, A. K. Rosen, P. A. Zrelak, B. Sadeghi, R. Baron, J. Cuny, H. M. Kaafarani, J. J. Geppert, and P. S. Romano. Designing an abstraction instrument: lessons from efforts to validate the ahrq patient safety indicators. *Jt Comm J Qual Patient Saf*, 37(1):20–8, 2011.

[13] G.K. Tayi and D.P. Ballou. Examining data quality. *Communications of the ACM*, 41(2):54–57, 1998.

[14] P. Oliveira, F. Rodrigues, P. Henriques, and H. Galhardas. A taxonomy of data quality problems. In *2nd Int. Workshop on Data and Information Quality*, pages 219–233, 2005.
[15] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey. Comorbidity measures for use with administrative data. *Med Care*, 36(1):8–27, 1998.

Alberto Freitas is an assistant professor at the Department of Health Information and Decision Sciences, Faculty of Medicine, University of Porto, Portugal, and researcher at CINTESIS—Center for Research in Health Technologies and Information Systems. In 2007, he obtained his PhD from University of Porto with a thesis on knowledge discovery in hospital data for management support. His research interests are focused on Health Informatics, Data Mining in Medicine, Cost-Sensitive Learning, Data Quality, and Performance and Quality Indicators.

Juliano Gaspar is PhD student in Health Informatics program at Women's Health, Faculty of Medicine of Federal University of Minas Gerais. Concluded a Masters in Medical Informatics at the Faculty of Medicine of University of Porto (2011), after obtaining a degree in Computer Science from the Universidade do Vale do Itaja (2006). He is a member of the Center for Health Informatics of Faculty of Medicine of the Federal University of Minas Gerais, where develops the activity of scientific research in healthcare and information technology. His researches are based on the Epidemiology through the Health Indicators, Health informatics and Maternal and Child Health.

Nuno Rocha is a BSc graduate, currently undertaking a MSc degree in ISEP School of Engineering at the Polytechnic of Porto (2012). He is a junior researcher at CINTESIS with a contribution in the development of Electronic Health Record (EHR) web applications, including data collection and analysis, through data mining techniques. His main projects are related to Obstetrics and Maternal care monitoring, Psychiatric patient records, Infectious Diseases and Anesthesiology records.

Goreti Marreiros is professor at the Polytechnic of Porto's Institute of Engineering (ISEP/IPP) and researcher at the Knowledge Engineering and Decision Support Research Group (GECAD). Her main areas of interest are Multi-Agent Systems, Emotional Agents, Persuasive Argumentation and Group Decision Support Systems. She received her PhD in informatics from the University of Minho.

Altamiro da Costa-Pereira, MD, obtained the degree of Doctor of Philosophy (PhD) from the University of Dundee, Scotland, in 1993. He is responsible for several under and post-graduate disciplines and courses, including a doctoral program in Clinical and Health Services Research, at the Faculty of Medicine, University of Porto. He undertook research projects in national and international institutions, in the fields of epidemiology, medical informatics and clinical research, publishing more than 300 scientific papers; he participated in more than 50 national and international panels and evaluating commissions of fellowships, projects and scientific research teams in the field of life and health sciences and technologies, being regularly invited by the European Commission as an expert in information technologies applied to healthcare. He is Director of the Department of Health Information and Decision Sciences, and also coordinates CINTESIS.