LINK TO ORIGINAL ARTICLE
LINK TO INITIAL CORRESPONDENCE

Reply to 'Mining electronic health records: an additional perspective'

Peter B. Jensen, Lars J. Jensen and Søren Brunak

We thank Hurdle *et al.* (Mining electronic health records: an additional perspective. *Nature Reviews Genetics* 18 Dec 2012 (doi:10.1038/nrg3208-c1)[1] for the appreciation of our Review on the new dimension that text mining brings to data integration within the health-care sector (Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* **13**, 395–405)[2]. Data integration across health registries has a long tradition that predates personal identification numbers. After such numbers were introduced, however, data integration became easier and more precise, leading to a situation in which entire countries could be regarded as cohorts for epidemiological investigation[3]. The Nordic countries were among the first to make such investigations possible.

The emphasis of our article was on electronic health records (EHRs), their potential and their current under-exploitation. Table 1 was never intended to be an exhaustive list of resources for health data integration. The Utah Population Database (UPDB) is indisputably a highly valuable resource that has been used to make seminal discoveries. However, in our understanding, it shares many similarities with existing national health registries, particularly in the Nordic countries[4]. That is, health information in the form of, for example, ICD-9 or ICD-10 codes from reimbursement data can be linked to data on birth, causes of death, socio-economic data and family relations. This can be done either through unique personal identification numbers, as in the Nordic registries, or using sophisticated record linkage techniques, as in UPDB. We do not question the importance of such resources in epidemiological studies, which is evidenced by papers referenced by Hurdle *et al.*[1] and countless other examples[5–7]. Rather, we focused on resources that more explicitly integrate EHRs: for example, the Vanderbilt BioVU and i2b2 resources[8,9].

Although we argue for common standards for interoperability and content models in the long term, we certainly do not see these as prerequisites for high-quality research. Indeed, our own work on medical text mining is based on data that do not meet such standards, and the analyses thus require extensive preprocessing and curation[10]. Resources such as the UPDB show what is possible with integration of heterogeneous data given sufficient effort and expertise. However, if in the future more data in EHR systems adhere to common standards, we believe a wider range of biomedical researchers will be able to carry out large-scale national and international integrative studies, incorporating detailed aspects of phenotypical and genetic data.

We thank Hurdle *et al.*[1] for giving us the additional opportunity to discuss how we think the existing excellent possibilities for carrying out epidemiological work on the basis of resources such as the UPDB and the Nordic health registries will be expanded by text mining of the unstructured data in EHRs. In particular, we believe that integration of such data with population-wide sequencing projects will make for an exiting future for our research field.

The authors are at the NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark.

Søren Brunak is also at the Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark.

Correspondence to S.B.
e-mail: brunak@cbs.dtu.dk

doi:10.1038/nrg3208-c2
Published online 18 December 2012

1. Hurdle, J, F., Smith, K. R. & Mineau, G.P. Mining electronic health records: an additional perspective. *Nature Rev. Genet.* 18 Dec 2012 (doi:10.1038/nrg3208-c1).
2. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Rev. Genet.* **13**, 395–405 (2012).
3. Frank, L. Epidemiology. When an entire country is a cohort. *Science* **287**, 2398–2399 (2000).
4. Thygesen, L. C., Daasnes, C., Thaulow, I. & Bronnum-Hansen, H. Introduction to Danish (nationwide) registers on health and social issues: structure, access, legislation, and archiving. *Scand. J. Publ. Health* **39**, 12–16 (2011).
5. Madsen, K. *et al.* A population-based study of measles, mumps and rubella vaccination and autism. *N. Engl. J. Med.* **347**, 1477–1482 (2002).
6. Sørensen, T. I., Nielsen, G. G., Andersen, P. K. & Teasdale, T. W. Genetic and environmental influences on premature death in adult adoptees. *N. Engl. J. Med.* **318**, 727–732 (1988).
7. Williams, R. R. Nature, nurture, and family predisposition. *N. Engl. J. Med.* **318**, 769–771 (1988).
8. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).
9. Murphy, S. N. *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.* **17**, 124–130 (2010).
10. Roque, F. S. *et al.* Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput. Biol.* **7**, e1002141 (2011).

Acknowledgements
The authors were supported in part by the Villum Kann Rasmussen Foundation, the Novo Nordisk Foundation and the Danish Research Council for Strategic Research.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Center for Biological Sequence Analysis — Technical University of Denmark: http://www.cbs.dtu.dk
The Novo Nordisk Foundation Center for Protein Research — University of Copenhagen: http://www.cpr.ku.dk

ALL LINKS ARE ACTIVE IN THE ONLINE PDF