LINK TO ORIGINAL ARTICLE
LINK TO AUTHOR'S REPLY

Mining electronic health records: an additional perspective

John F. Hurdle, Ken R. Smith and Geraldine P. Mineau

We read with great interest the article by Jensen *et al.* (Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* **13**, 395–405)[1]. This was a well-written Review that summarized a large, complex and topical subject. To augment the article, and in particular to augment Table 1, we would like to point out that one of the earliest and most successful research databases that integrated diverse data sources with electronic health records (EHRs) is the Utah Population Database (UPDB) at the University of Utah, USA. Its earliest success was the identification of families with a high incidence of breast cancer; this research led to the discovery of the breast cancer genes *BRCA1* and *BRCA2* (REFS 2–5). A crucial component of the UPDB — one that allowed it to probe genetic inheritance long before gene sequencing was widely available — was the linking of diverse data sources with family pedigrees that were originally supplied by the Utah Genealogical Society and that were later updated by probabilistic matching with vital records from the Utah Department of Health (records such as birth, death and marriage certificates).

As Jensen *et al.*[1] pointed out, there is much to be gained by mining data in EHRs, especially when they are linked to other sources. Using the UPDB, researchers at the University of Utah have made discoveries across a wide variety of disciplines in addition to oncology, including gynaecology[6], autoimmune disease[7], spinal abnormalities[8], ophthalmology[9], gastroenterology[10] and gerontology[11]. The utility of the UPDB derives from the integration of the EHRs from two large health-care networks in the state of Utah[12,13] coupled with high-quality data from the Utah Department of Health, the Utah and Idaho cancer registries and a deep, expansive family pedigree database.

Finally, we would qualify the conclusion in Jensen *et al.*[1], which stated that "True data interoperability requires the development and implementation of standards and clinical-content models for the unambiguous representation and exchange of clinical meaning". Like those authors, we firmly advocate the wide adoption of standards, even for clinical-content models. However, the success of the UPDB, built as it was on well-crafted probabilistic matching, serves as an example of how quality research can be conducted even in the absence of uniform data standards.

John F. Hurdle is at the University of Utah School of Medicine, Health Sciences Education Building, Biomedical Informatics, 26 South 2000 East, Salt Lake City, Utah 84112, USA.

Ken R. Smith and Geraldine P. Mineau are at the University of Utah Huntsman Cancer Institute, 2000 Circle of Hope, Salt Lake City, Utah 84112, USA.

Correspondence to J.F.H.
e-mail: john.hurdle@utah.edu

doi:10.1038/nrg3208-c1
Published online 18 December 2012

1. 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).
2. Goldgar, D. *et al.* Chromosome 17q linkage studies of 18 Utah breast cancer kindreds. *Am. J. Hum. Genet.* **52**, 743 (1993).
3. Goldgar, D. E. *et al.* A large kindred with 17q-linked breast and ovarian cancer: genetic, phenotypic, and genealogical analysis. *J. Natl Cancer Institute* **86**, 200–209 (1994).
4. Neuhausen, S. L. *et al.* Haplotype and phenotype analysis of six recurrent *BRCA1* mutations in 61 families: results of an international study. *Am. J. Hum. Genet.* **58**, 271 (1996).
5. Tavtigian, S. *et al.* The complete *BRCA2* gene and mutations in chromosome 13q-linked kindreds. *Nature Genet.* **12**, 333–337 (1996).
6. Allen-Brady, K. *et al.* Identification of six loci associated with pelvic organ prolapse using genome-wide association analysis. *Obstetr. Gynecol.* **118**, 1345 (2011).
7. Frech, T. *et al.* Heritability of vasculopathy, autoimmune disease, and fibrosis in systemic sclerosis: a population-based study. *Arthritis Rheum.* **62**, 2109–2116 (2010).
8. Patel, A. A., Spiker, W. R., Daubs, M., Brodke, D. S. & Cannon-Albright, L. A. Evidence of an inherited predisposition for cervical spondylotic myelopathy. *Spine* **37**, 26 (2012).
9. Wang, X. *et al.* Using the Utah Population Database to assess familial risk of primary open angle glaucoma. *Vision Res.* **50**, 2391–2395 (2010).
10. Guthery, S. L., Mineau, G., Pimentel, R., Williams, M. S. & Kerber, R. A. Inflammatory bowel disease aggregation in Utah kindreds. *Inflamm. Bowel Dis.* **17**, 823–830 (2011).
11. Kerber, R. A., O'Brien, E., Boucher, K. M., Smith, K. R. & Cawthon, R. M. A. Genome-wide study replicates linkage of 3p22-24 to extreme longevity in humans and identifies possible additional loci. *PLoS ONE* **7**, e34746 (2012).
12. DuVall, S. L., Fraser, A. M., Rowe, K., Thomas, A. & Mineau, G. P. Evaluation of record linkage between a large healthcare provider and the Utah Population Database. *J. Am. Med. Inform. Assoc.* **19**, e54–e59 (2011).
13. DuVall, S., Fraser, A., Kerber, R., Mineau, G. & Thomas, A. The impact of a growing minority population on identification of duplicate records in an enterprise data warehouse. *Stud. Health Technol. Inform.* **160**, 1122 (2010).

Competing interests statement
The authors declare no competing financial interests.