Identifying correlations in electronic patient records

August 25 2011

A new study demonstrates how text mining of electronic health records can be used to create medical term profiles of patients, which can be used both to identify co-occurrence of diseases and to cluster patients into groups with highly similar clinical features. The study, carried out in Denmark by a multi-disciplinary group of bioinformaticians, systems biologists and clinicians, will be published in the open-access journal *PLoS Computational Biology* on 25th August 2011.

Health records contain detailed phenotypic information on the clinical profile of each individual patient; however, a large part of the clinical features are described in free text produced by hospital staff often covering many years of hospitalization.

"Using our text mining approach on the free text in the records, we identified roughly ten times as many medical terms characterizing each patient as were manually included by the hospital staff. Worldwide, the manually inserted medical terms in medical records are heavily biased by local practice and billing purposes. Using our method we obtained a much more fine-grained clinical characterization of each patient, which ultimately also may be very valuable for choosing personalized treatment regimes", says Professor Søren Brunak from the Technical University of Denmark and the University of Copenhagen who led the team behind the research project.

The team used the "International Classification of Disease" terminology, maintained by the WHO as a controlled vocabulary, as the basis for the

analysis. "The fact that terminologies like ICD have been translated word by word between languages makes it possible in principle to use the same term profiles across language barriers and combine cohorts across countries" says author Professor Lars Juhl Jensen from the University of Copenhagen.

The research group identified a large number of diseases and symptoms which co-occur much more than expected when compared to the individual frequencies of the diseases. The group subsequently mapped these correlations to the genetic level by investigating gene overlaps in protein interaction networks already linked to the individual diseases. "The aim here is to discover a possible genetic cause behind the disease correlations observed, thus interfacing the electronic patient record data directly to the DNA sequencing of human individuals", says Brunak.

More information: Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, et al. (2011) Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. PLoS Comput Biol 7(8): e1002141. doi:10.1371/journal.pcbi.1002141