Unlocking the Potential of Electronic Health Records for Translational Research
Findings from the Section on Bioinformatics and Translational Informatics

Y. L. Yip, Section Editor for the IMIA Yearbook Section on Bioinformatics and Translational Informatics
Knowledge Management, Merck Serono International S.A., 9 Chemin des Mines, 1202 Geneva, Switzerland

Summary

Objectives: To review current excellent research and trend in the field of bioinformatics and translational informatics with direct application in the medical domain.
Method: Synopsis of the articles selected for the IMIA Yearbook 2012.
Results: Six excellent articles were selected in this Yearbook's section on Bioinformatics and Translational Informatics. They exemplify current key advances in the use of patient information for translational research and health surveillance. First, two proof-of-concept studies demonstrated the cross-institutional and -geographic use of Electronic Health Records (EHR) for clinical trial subjects identification and drug safety signals detection. These reports pave ways to global large-scale population monitoring. Second, there is further evidence on the importance of coupling phenotypic information in EHR with genotypic information (either in biobank or in gene association studies) for new biomedical knowledge discovery. Third, patient data gathered via social media and self-reporting was found to be comparable to existent data and less labor intensive. This alternative means could potentially overcome data collection challenge in cohort and prospective studies. Finally, it can be noted that metagenomic studies are gaining momentum in bioinformatics and system-level analysis of human microbiome sheds important light on certain human diseases.
Conclusions: The current literature showed that the traditional bench to bedside translational research is increasing being complemented by the reverse approach, in which bedside information can be used to provide novel biomedical insights.

Keywords

Medical informatics, International Medical Informatics Association, yearbook, bioinformatics, translational informatics, patient data, electronic health record, metagenomics

Introduction

Translational research is viewed by many people as the science of taking research 'bench-to-bedside'. As already highlighted in Yearbook 2008 [1], one of the foci in translational informatics has been on the semantic integration of heterogeneous sources of data. In the same year, an article published in Nature described 'The Full Cycle' of translational research, and advocated the importance of the reverse 'bedside to bench' approach [2]. While the topic was on the beneficial use of clinical trials data to inspire new avenue of research, the concept can be broadened to include the general use of patient information to advance research. A survey of recent articles showed that EHR systems, either alone or integrated with other biomedical resources, constitute a rich resource for many areas of data-driven knowledge discovery. In particular, phenotypic information collected in EHR could be used to uncover genetic variants influencing treatment regime, provide new insights on gene-disease associations, and also refine *in silico* models for clinical adaptation [3-6]. Recent reports on the incorporation of genetic information in a standards-based EHR system prototype, as well as frameworks providing access to cross-institutional, geographically separated EHR systems further promise the important role EHR may play in reverse translational research [7-9].

In the bioinformatics field, metagenomic studies are gaining momentum and the enormous influence that the microbiome exerts on the body's response to the environment and the pathogenesis of disease is increasingly being recognized and studied at a system-level [10]. The microbiome represents another aspect of environmental factors already highlighted in last year's Yearbook [11].

Best Paper Selection

The best paper selection of articles for the section Bioinformatics and Translational Informatics in the IMIA Yearbook 2012 follows the tradition of previous yearbooks and expands the scope by including topics covering translational informatics. The inclusion is timely as translational informatics could be regarded as a formidable bridge between bioinformatics and medical informatics. Indeed, while there is a clear distinction between the two fields several years ago, the boundary is becoming less notable, with papers in one domain appearing in journals normally reserved for the other domain and vice versa. As a result of the review process, six articles were selected from international peer-reviewed journals in the fields of biomedicine, medical informatics, translational informatics and bioinformatics [3.4, 8-10, 13].

The first four papers described the use of electronic health records in different research settings. Anderson *et al.* described the Cross-Institutional Clini-

cal Translational Research (CICTR) project which aimed to identify potential subjects for clinical trials from multiple medical centers through a federated query tool based on the Integrating Biology and the Bedside (i2b2) platform [8]. The i2b2 is a scalable informatics framework that aims to enable researchers to use clinical data for discovery research [12]. In another larger cross-institutional and cross-country study, Coloma *et al*. reported the successful combination of eight EHR databases for drug adverse event detection. Both studies pave way to large-scale, cross-institutional and global patient identification or monitoring system [9]. In parallel, two papers showed how patient information in EHR can be used to advance biomedical knowledge and derive better therapeutic strategy. Davis and Chawla exploited patient phenotypic data stored in EHR to complement gene association studies for the modeling of disease co-morbidities using an integrated network approach [4]. Birdwell *et al*., on the other hand, provided further evidence on the usefulness of coupling DNA biobank to EHR to uncover novel findings, i.e. new pharmacogenomic predictors of tacrolimus dose requirement in kidney transplantation [3]. Besides the use of EHR for structured patient information, Weitzmann *et al*. leveraged on the power of social media and its increasing popularity to pioneer a new way for public health surveillance [13]. A 'facebook-like' software application was provided to an online international diabetes community to allow users to self-report glycemic control. Finally, Greenblum *et al*. presented a system-level metagenomic approach to study the human microbiome, its organization, and its impact on obesity and inflammatory bowel disease [10]. Their findings demonstrate the increasing importance of microbiome and host interactions in human health.

Table 1 presents the selected papers. A brief summary of the selected best papers can be found in the appendix of this report.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2010 in the section 'Bioinformatics and Translational Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Bioinformatics and Translational Informatics

- Anderson N, Abend A, Mandel A, Geraghty E, Gabriel D, Wynden R, Kamerick M, Anderson K, Rainwater J, Tarczy-Hornoch P. Implementation of a deidentified federated data network for population-based cohort discovery. J Am Med Inform Assoc 2011; Epub
- Birdwell KA, Grady B, Choi L, Xu H, Bian A, Denny JC, Jian M, Vranic G, Basford M, Cowan JD, Richardson DM, Robinson MP, Ikizler TA, Ritchie MD, Stein CM, Haas DW. The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement in kidney transplant recipients. Pharmacogenet Genomics 2012;22:32-42.
- Coloma PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, Mazzaglia G, Giaquinto C, Corrao G, Pedersen L, Van der Lei J, Sturkenboom M. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf 2011;20:1-11.
- Davis DA, Chawla NV. Exploring and exploiting disease interactions from multi-relational gene and phenotype network. PLoS One 2011;6(7):e22670
- Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome revieals topological shifts associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci USA 2012;109(2):594-9.
- Weitzmann ER, Adida B, Kelemen S, Mandl KD. Sharing data for public health research by members of an international online diabetes social network. PLoS One 2011; 6(4): e19256

Conclusions and Outlook

As the fields of bioinformatics and medical informatics further converge, bidirectional information flow between biomedical research and patient data will be made ever easier to enable the full cycle of translational research. In the coming years, several current trends on the use of patient information for research and surveillance purposes are likely to continue: (1) the increased coupling of the EHR and personal genomic data to advance personalized medicine; (2) cross-institutional and cross-countries use of different EHR databases to enable large-scale clinical studies; (3) the use of social media to collect patient self-reporting data for disease monitoring. Research potential enabled by these new avenues could be enormous.

However, as highlighted in several papers, issues or bottlenecks related to the wide-spread use or sharing of patient or clinical information are no longer in the technical realm, but rather social. Considerable institutional co-ordination, consensus building and end-user engagement are needed to realize the full potential of the resources. Simi-larly, ethical issues related to anonymity and privacy have to be adequately addressed.

References

1. Yip YL. The promise of systems biology in clinical applications. Findings from the Yearbook 2008 Section in Bioinformatics. Yearb Med Inform 2008;102-4.
2. Ledford H. The full cycle. Nature 2008;453:843-5.
3. Birdwell KA, Birdwell KA, Grady B, Choi L, Xu H, Bian A, et al. The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement in kidney transplant recipients. Pharmacogenet Genomics 2012;22:32-42.
4. Davis DA, Chawla NV. Exploring and exploiting disease interactions from multi-relational gene and phenotype network. PLoS One 2011;6(7): e22670
5. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol 2011; 7(8): e1002141
6. Stamatakos GS, Georgiadi EC, Graf N, Kolokotroni EA, Dionysiou DD. Exploiting clinical trial data drastically narrows the window of possible solutions

to the problem of clinical adaption of a multiscale cancer model. PLoS One 2011;6(3): e17594

7. Jing X, Kay S, Marley T, Hardiker NR, Cimino JJ. Incorporating personalized gene sequence variants, molecular genetics knowledge, and health knowledge into an EHR prototype based on the continuity of care record standard. J Biomed Inform 2012;45(1):82-92.

8. Anderson N, Abend A, Mandel A, Geraghty E, Gabriel D, Wynden R, et al. Implementation of a deidentified federated data network for population-based cohort discovery. J Am Med Inform Assoc 2011; Epub

9. Coloma PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf 2011;20:1-11.

10. Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveials topological shifts associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci USA 2012;109(2): 594-9.

11. Yip YL. Genome, and beyond. Findings from the Yearbook 2011 Section in Bioinformatics. Yearb Med Inform 2011;6(1):156-9.

12. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012;19(2):181-5.

13. Weitzmann ER, Adida B, Kelemen S, Mandl KD. Sharing data for public health research by members of an international online diabetes social network. PLoS One 2011;6(4):e19256.

Correspondence to:
Dr. Yum Lina Yip
Knowledge Management
Merck Serono S.A.
9 Chemin des Mines, Geneva, Switzerland
E-mail: lina.yip.sonderegger@merckgroup.com

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2012, Section Bioinformatics and Translational Informatics*

Anderson N, Abend A, Mandel A, Geraghty E, Gabriel D, Wynden R, Kamerick M, Anderson K, Rainwater J, Tarczy-Hornoch P

---

* The complete papers can be accessed in the Yearbook's full electronic version, provided that the article is freely accesible or that your institution has access to the respective journal.

Implementation of a deidentified federated data network for population-based cohort discovery

JAMIA 2011; Epub

This article introduced the Cross-Institutional Clinical Translational Research (CICTR) project which sought to pilot approaches to identify potential subjects for clinical trial recruitment across multiple, geographically separated integrated data repositories (IDR) systems in different medical centers. The CICTR project had two primary outcomes: (1) the implementation of a federated query tool based on the Integrating Biology and the Bedside (i2b2) platform on deidentified data for over 5 million patients and 5 data sources across 3 independent institutional IDRs; (2) the development and the testing of an iterative process model based on partnerships building, system requirements, technical architecture, and evaluation/promotion to manage and coordinate new use cases and data sources. The authors concluded that coordination across diverse clinical and research environments to align technical, semantic, and policy issues remained predominantly a social challenge. The continued sustainability of the pilot network would require strong partnership with scientific end users at local sites, and considerable consensus-building and interdisciplinary education. At the time of the article submission, the projects had engaged with additional Clinical and Translational Science Awards (CTSA) partners.

Birdwell KA, Grady B, Choi L, Xu H, Bian A, Denny JC, Jian M, Vranic G, Basford M, Cowan JD, Richardson DM, Robinson MP, Ikizler TA, Ritchie MD, Stein CM, Haas DW

The use of a DNA biobank linked to electronic medical records to characterize pharmaco-genomic predictors of tacrolimus dose requirement in kidney transplant recipients

Pharmacogenet Genomics 2012;22:32-42

There is a marked interindividual pharmacokinetic variability and a nar- row therapeutic index for the use of tacrolimus, an immunosuppressive calcineurin inhibitors widely prescribed to kidney transplant recipients. In this article, the authors investigated the use of DNA biobank and electronic medical record resources to identify drug absorption, distribution, metabolism, and elimination (ADME) gene variants associated with tacrolimus dose requirement. In particular, BioVU, the DNA biobank for Vanderbilt University was used in conjunction with the Synthetic Derivative, a deidentified version of Vanderbilt's electronic medical record, to identify initial candidates who met screening criteria. Broad ADME genotyping was subsequently performed on 446 kidney transplant recipients. Using this new approach, the authors were able to replicate the association of tacrolimus blood concentration to dose ratio with CYP3A5 rs776746, and further identified associations with nine variants in linkage disequilibrium with rs776746, including eight CYP3A4 variants. The authors concluded that the use of a DNA biobank coupled with clinical information could be a reliable alternative to more time and labor intensive cohort studies, and the ability to link genotype and phenotype through BioVU/Synthetic Derivative made it a powerful platform for genetic association studies.

Coloma PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, Mazzaglia G, Giaquinto C, Corrao G, Pedersen L, Van der Lei J, Sturkenboom M

Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project

Pharmacoepidemiol Drug Saf 2011;20:1-11

The EU-ADR Project is a European project launched in 2008 that aims to exploit information from various electronic healthcare record (EHR) databases in Europe to produce a computerized integrated system for the early detection of drug safety signals. In this proof-of-concept study, the au-

thors developed and tested a methodology that enables combining data from eight EHR databases of four countries (Denmark, Italy, Netherlands, the UK) using a distributed network approach. The common data framework allowed the data extraction to be done locally so that databases retained ownership of their respective data. Only the aggregated, de-identified data were shared with the rest of the network. The known association of NSAIDs and UGIB was used to demonstrate the sensitivity of the system by comparing the incidence rates of upper gastrointestinal bleeding (UGIB) and nonsteroidal anti-inflammatory drug (NSAID) utilization patterns. 39,967 incidence cases of UGIB were identified during the study period with a study population of 19,647,445 individuals corresponding to 59,929,690 person-year of follow-up. A statistically significant association between the use of any NSAID and increased risk for UGIB was confirmed in all databases. This important study proves that combining data from EHR databases of different countries to identify drug-adverse event associations is feasible. It also paves the way for a global-scale, EHR-based drug safety monitoring system.

Davis DA, Chawla NV

Exploring and exploiting disease interactions from multi-relational gene and phenotype network

PLoS One 2011;6(7):e22670

Disease mechanisms form a complex system, in which distinct factors such as genetic, environment and lifestyles come to play and have varying impact on problems such as co-morbidity and drug efficiency. In this study, the authors used patient medical history (phenotype data) spanning over 12 years and previously discovered disease-gene associations to construct, analyze, and compare disease interaction networks. By exploring both individual and combined interactions among these 2 lev-

els of disease data, novel insight into the interplay between genetics and clinical observations was obtained. Compared to previous studies, this approach merged heterogeneous data into a multi-relational network and thus offered a composite view. This study further demonstrated how the multi-relational structure could be applied to enhance the link prediction method of determining good targets for further gene association research, thereby leading to improved biological knowledge and clinical standards.

Greenblum S, Turnbaugh PJ, Borenstein E

Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease

Proc Natl Acad Sci USA 2012;109(2):594-9

The human microbiome (microbial communities populating the human body) plays a key role in human health because of its role in development, immunity and nutrition. In this study, the authors introduced a metagenomic systems biology computational framework, by integrating metagenomic data of the human gut microbiome with *in silico* systems-level analysis of metabolic networks. The work focused on the topology of this metagenome-based network and on the relationship between its topology and the host state. Briefly, fecal metagenomic data from 124 unrelated individuals, as well as six monozygotic twin pairs and their mothers were analyzed and community-level metabolic networks of the microbiome were generated. Both gene-level and network-level topological differences associated with obesity and inflammatory bowel disease (IBD) were identified. The authors suggested that lean and obese microbiomes differed primarily in their interface with the host metabolism, and further linked topological variations to community species composition. This metagenomic systems biology approach goes beyond tra-

ditional metagenomic comparative analysis and gene-centric view, and provides a complementary viewpoint so that valuable insight concerning the function of the microbiome as a system and its impact on the community and the host can be gained.

Weitzmann ER, Adida B, Kelemen S, Mandl KD

Sharing data for public health research by members of an international online diabetes social network

PLoS One 2011;6(4):e19256

Diabetes is a global health threat with an evolving disease morphology. Intensive population-wide monitoring and longitudinal tracking of diabetes are important for course correcting the disease. In this study, the authors tested a model for engaging an online international diabetes community (TuDiabetes) to share data for public health research and surveillance. They launched a social networking (SN) software application (TuAnalyze) with a "Facebook-like" environment to enable users to self-report glycemic control (Hemoglobin A1c or A1c value) and share disease information. There was substantial early adoption of the application with participation by 10-17% in the initial study period. 83.1% of the most recent A1c values reported were "current" (obtained within the past 90 days). There was also a high willingness to share, with 81.4% of users permitting data donation to the community display, and 34.1% of users displaying their A1cs on their SN profile page. Importantly, it was also found that the unadjusted aggregate A1c reported by US users was not different from the unadjusted aggregate of the 2007–2008 NHANES estimates. The authors concluded that online SNs may serve as efficient platforms for bidirectional communication with disease populations, and represent a significant extension of current reporting practice and capacity.