# Summary of Best Papers Selected for the IMIA Yearbook 2016, Section Clinical Research Informatics

### Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, Sweeney TE, Gyang E, Shah NH

### Learning statistical models of phenotypes using noisy labeled training data

### J Am Med Inform Assoc 2016;23(6):1166-73

Machine learning approaches running on real-world data are limited by the paucity of labeled training datasets. Traditionally, patient groups sharing a given phenotype are selected through rule-based definitions which creation and validation are time-consuming. This paper addresses the limitation of the generation of clinical phenotype descriptions and demonstrates the feasibility of utilizing semi-automatically labeled training sets to create phenotype models via machine learning, using a comprehensive representation of the patient medical record. The authors provide an extended background about i) manual rule-based definition of phenotypes for research purposes; ii) learning techniques (based on Natural Language Processing and/or other techniques) using manually created training sets (labeled cases and controls built from chart review), and iii) noise tolerant learning techniques.

Based on the phenotype definitions provided by the eMERGE [1] and OMOP [4] initiatives, the authors automatically identified within the Stanford Clinical Data Repository 32,581 possible cases for T2DM and 36,858 possible cases for MI. Using the Halpern *et al.* method based on "anchor" terms [2], they defined a list of keywords specific to the phenotypes of interest to semi-automatically generate noisy labeled training data. Then, a sample of 1,500 patient records - 750 patient records for each phenotype having a "noisy" label for the phenotype and 750 controls taken in the extract disjoint with possible cases (silver standard) - was used to train the XPRESS (eXtraction of Phenotypes from Records using Silver Standards) model. The building of XPRESS models consisted of feature engineering from structured and unstructured data and learning statistical models from the noisy labeled data. The performance of the models was evaluated against a gold standard consisting of a clinician-reviewed evaluation set (gold standard: cases and controls created by five clinicians, disjoint from the records used for the training). The models for T2DM and MI achieved a precision and accuracy of 0.90, 0.89, and 0.86, 0.89, respectively. Local implementations of the previously validated rule-based definitions for T2DM and MI achieved precision and accuracy of 0.96, 0.92 and 0.84, 0.87, respectively. The authors demonstrated that they can learn phenotype models of chronic and acute phenotypes from 4,135 noisy labeled training samples (XPRESS models) acquired at a negligible cost with the same performance as from 2,026 manually labeled, zero-error training samples. Using imperfectly labeled data, the method provides an alternative to manual labeling for creating training sets. Such an approach may be used to create local phenotype models and can accelerate research with large observational healthcare datasets. Further research in feature engineering and in the specification of the keyword list can improve the performance of the models and the scalability of the approach.

### Pfiffner PB, Pinyol I, Natter MD, Mandl KD

### C3-PRO: Connecting ResearchKit to the Health System Using i2b2 and FHIR

### PloS One 2016;11(3):e0152722

As new mobile technologies are more widely adopted, their use for care and research is being more and more efficient. One of the actual challenges is to connect research Apps to the healthcare system and use real life patient-generated data in order to improve pharmacovigilance and to obtain medication observance data for post market studies or other usages. In March 2015, Apple Inc. deployed a new open source framework to help research promoters to build easy smartphones Apps for clinical studies. To complete the system, the authors extended the Apple ResearchKit with a Consent, Contact, and Community framework for Patient Reported Outcomes (namely C3-PRO). The aim of this extension is to connect the ResearchKit App to the widely used clinical research IT infrastructure i2b2. C3-PRO enables a method to create eligibility criteria question, informed consent, and participant surveys using FHIR data formats. Data is encrypted prior to be sent over the Internet. It is then pushed into an i2b2 FHIR compatible cell. The paper describes the complete system and the data flows including security measures both in terms of data processing and at the App level. The system can collect data anonymously, using the UUID (Universally Unique Identifier) of the device as identifier. The system can also capture sensor-based data. Using the system, recruitment for studies can be done more widely and faster. The resulting data processing is then taking advantage of i2b2 generic architecture to process classic statistics and produce first reports. Besides, authors are working on mechanisms for data-linkage with existing cohorts as well as a cross platform version or their kit (Android/Iphone). By leveraging the FHIR formats, C3-PRO enables survey question and consent libraries to become standardized and used across studies.

### Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B

### The FAIR Guiding Principles for scientific data management and stewardship

### Sci Data 2016 Mar 15;3:160018

The current digital ecosystem of scholarly data publication still prevents us from extracting the maximum benefit from research

investments. Science funders, publishers, and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Main barriers to data reusability are not technical. An appropriate set of basic principles to data stewardship to be followed by database owners, data managers, or data scientists is proposed in this paper to integrate and propagate digital object "best design" rules. The authors present four foundational principles, the FAIR guiding principles, that are related but independent and separable. The FAIR guiding principles are setting basic rules so that data should be: Findable, Accessible, Interoperable and Reusable. Operational rules are defined for each principle. For instance, to be *Findable*, a unique and persistent identifier should be assigned to data and metadata, rich metadata should describe data, and metadata and data should be registered and indexed in a searchable source. To be *Accessible*, (meta) data are retrievable by their identifier using a standard, open, and free protocol allowing authentication and authorization procedures, and allowing metadata to be accessible even when data is not. To be *Interoperable*, (meta)data should use a formal, accessible, and shared set of broadly applicable languages and vocabularies for knowledge representation. And finally, to be *Reusable* (meta) data should be richly described with a plurality of attributes and be released with provenance and clear licensing information. These principles do not suggest for any specific technology, nor standard or implementation-solution. Many scientific datasets or projects, such as Dataverse, FAIRDOM, ISA, Open PHACTS, or UniProt, are already implementing some of these principles. Although FAIR principles are not a technical standard, they put a specific emphasis on enhancing the ability of machines to automatically find and use data, in addition to supporting its reuse by individuals.

### Harmanci A, Gerstein M

### Quantification of private information leakage from phenotype-genotype data: linking attacks

### Nat Methods 2016 Mar;13(3):251-6

As the number and size of phenotype and genotype datasets increase, the privacy protection of individuals is emerging as an important issue. This paper focuses on the evaluation of the risk of privacy breaches in releasing genomics datasets. Harmanci *et al.* investigated how far, molecular phenotype data (such as gene expression level) can be - in contrast to DNA variants - considered as free of identifying information as it is generally assumed.

The authors provide a background about the growing list of quasi identifiers in molecular phenotype datasets and about two different types of privacy breaches. These privacy breaches result from either detecting whether an individual with known genome has participated to a study or cross-referencing of multiple seemingly independent genotype and phenotype datasets (knowing that the number of potentially linkable datasets will increase). They propose a framework for practical instantiation of linking attacks using a genotype dataset and publicly available anonymized phenotype datasets and genotype-phenotype correlations. The authors emphasize the need of statistical quantification methods to objectively quantify the risk of linking attacks before releasing a genotype dataset. They propose two measures: the cumulative individual characterization information (ICI) and the genotype predictability. ICI is described as the total amount of information in a set of variant genotypes that can be used in a linking attack. For a set of variants, genotype predictability measures how predictable genotypes are, given the gene expression levels. A three-step framework for instantiating linking attacks is presented. Based on the framework implementation on a test set, authors demonstrated that more than 95% of individuals are vulnerable and they observed that the extremity attacks (extreme of the gene expression levels observed with extremes of the phenotypes) can link family members within the dataset. Once the risk assessment is performed, several strategies can be set to minimize risks. For example k-anonymization proposes to censor entries or adding noise into the dataset on specific data points that have been characterized as possible leaks (ICI). Finally, inclusion of biological utility measures should be done along with the risk assessment. The methods proposed by the authors can be integrated directly into the existing risk assessment and management strategies.