Appendix: Summary of Best Papers Selected for the 2019 IMIA Yearbook, Section Clinical Research Informatics

Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W

Federated learning of predictive models from federated Electronic Health Records

Int J Med Inform 2018 Apr;112:59-67

Centralized machine learning methods are typically used to train predictive models from data that are aggregated into large central repositories. This paper describes an alternative machine learning method applicable to massive data residing in different locations and owned by different entities that could not be aggregated into a single repository due to technical and/or privacy concerns. Brisimi *et al.* developed a new federated algorithm - the cluster Primal Dual Splitting (cPDS) algorithm - for solving the large-scale sparse Support Vector Machine (sSVM) problem in a decentralized fashion. They applied this new algorithm to a dataset of de-identified Electronic Healthcare Records (EHRs) from the Boston Medical Center for predicting heart failure hospital admissions based on patients' medical history described in their distributed EHRs. The federated optimization scheme cPDS enables multiple data holders to collaborate and converge to a common predictive model, without explicitly exchanging raw data. This distributed algorithm accurately differentiates between patients that are likely or unlikely to be hospitalized within a target year. With a prediction accuracy of 0,7806 measured by the Area Under the Receiver Operating Characteristic Curve (AUC), the cPDS performs better than the other methods used to solve the sSVM problem. The authors demonstrate that cPDS converges faster than both centralized methods and alternative distributed algorithm. Important features that are predictive of future hospitalizations have been discovered by the algorithm, such as age, diagnosis of heart failure in the year before the target year, admission due to heart failure or other circulatory system diagnoses one year before the target year, thus providing a way to interpret the classification results and inform prevention efforts. At a time of increasing preference for distributed querying to avoid centralizing data, this paper makes a great contribution by providing a validated approach to distributed machine learning for such architectures.

Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N

Initializing a hospital-wide data quality program. The AP-HP experience

Comput Methods Programs Biomed 2018 Nov 9

There is increasing recognition of the importance of assessing and improving the data quality (DQ) of hospital EHRs, not only for clinical care purposes but to enable robust clinical research inferences from the data. This is of of the two 2018 best papers selected on DQ for this section. Unlike many publications that focus only on assessment, this paper has been selected because it tackles the challenge holistically examining how DQ can be improved, and undertook the research across 37 hospitals in the Paris region (the Assistance Publique – Hôpitaux de Paris, AP-HP). Daniel *et al,.* designed and conducted DQ campaigns consisting of five phases: defining the scope, measuring, analyzing, improving, and controlling DQ. They applied this in two domains - patient identification and healthcare services. Through EHR data profiling across the AP-HP network, comprising a repository of 8.8 million patients, the authors identified 11 data quality issues. These were categorized into completeness, conformance, and plausibility DQ issues. The root causes of these issues were found to be errors from data originators, ETL issues, or limitations of the source EHR data, and these insights informed a DQ improvement campaign. The improvement strategies targeted staff communication and teaching (leaflets, videos, feedback), the engagement of patient registration staff and health professionals (DQ campaigns, updating specialty vocabulary tables), patient engagement (in rechecking their information), and information system improvements such as computerised DQ checks and fixing record merger errors. These action plans, though only partially implemented at the time of publication, resulted in significant improvement of DQ measures. This research was included as a best paper because it provides insights into the actual data quality observed across 37 hospitals, linked to the kinds of campaign actions that can be implemented: the research goes beyond assessment to improvement.

Estiri H, Stephens KA, Klann JG, Murphy SN

Exploring completeness in clinical data research networks with DQe-c

J Am Med Inform Assoc 2018 Jan 1;25(1):17-24

A new generation of clinical research platforms offers the capability to reuse on a large (multi-site, federated) scale routinely collected hospital EHR data for clinical research, such as clinical trials and big data mining. Since data quality (DQ) imperatives to support continuity of care and to support reuse for research are quite different, DQ poses important concerns for secondary use of EHR data and remains a challenge for research data networks using non-scalable ad hoc solutions. Estiri's paper is one of the two 2018 best papers selected on DQ for this section. The authors developed an open source, interoperable, and scalable DQ assessment tool able to measure the completeness and conformance of data items within an EHR or CDW data model. They describe the iterative implementation of a web-based tool - DQe-c - across different institutions focusing on interoperability and scalability to large databases. The DQe-c has been evaluated on a sample dataset of 200,000 randomly selected patient records with an encounter since January 1, 2010, extracted from the Research Patient Data Registry at Partners HealthCare. The web-based report produced by DQe-c is organized into four sections: load and test details (list of the tables of the EHR or CDW's common data model (CDM) and table-level size and completeness), completeness test (missing data for each column of the tables), data model conformance test (rate of orphan records based on the CDM), and fitness for purpose test (missingness in key clinical indicators such as ethnicity

data or blood pressure for example). This best paper contributes to the body of open algorithms and tools to permit comparable DQ assessments which can be run on different architectures (EHRs, PCORnet, OMOP) paving the way to systematic evaluation of DQ across distributed networks.

Sylvestre E, Bouzillé G, Chazard E, His-Mahier C, Riou C, Cuggia M

Combining information from a clinical data warehouse and a pharmaceutical database to generate a framework to detect comorbidities in electronic health records

BMC Med Inform Decis Mak 2018 Jan 24;18(1):9

Comorbidities are an increasing healthcare challenge and are important for accurate clinical research. Electronic health records, and clinical data warehouses derived from them, tend to be incomplete with respect to comorbidities. This may be because healthcare organisations are more likely to capture coded diagnoses that are relevant to the services they are providing. However, clinicians are normally keen to ensure they have a complete and up-to-date medication list. This best paper research assessed whether it is possible to use medication lists to identify missing comorbidities, to make a clinical data warehouse more complete and accurate for research use. Sylvestre et al., from the Rennes University Hospital, developed an algorithm that analyses medication lists to identify clinical indications that were not already documented as co-morbidities. For accuracy, the authors only included drug prescriptions with precise indications, by combining data from the French Comorbidity List with the French Theriaque drug database. They additionally selected laboratory tests with very precise indications to identify health conditions that the requesting clinicians must have known about, but which were not listed as coded comorbidities. Their analysis included 4,312 hospital stays with a qualifying prescription. Among the 4,312 patients of the general dataset, 68.4% had at least one drug prescription without a corresponding ICD-10 code. The comorbidity diagnoses suggested by the algorithm were confirmed by experts in 20.3% of reviewed cases. A specialized extract of 122 Ear Nose and Throat hospital stays was used to further evaluate the algorithm, within which comorbidity diagnoses suggested by the algorithm were confirmed in 44.6% of the cases. This research was included as a best paper because it offers and validates a relatively simple approach to semantic enrichment of CDWs, that could be replicated almost anywhere. It highlights the importance of co-morbidity coding, and of poor current practice in such coding.